DISERTASI SS143506

IDENTIFIKASI KESAMAAN POLA DOKUMEN TEKS BERDASARKAN KEMUNCULAN *TERM* DALAM KALIMAT

SOEHARDJOEPRI
NRP. 1310301002

PROMOTOR/CO-PROMOTOR
Prof. Drs. NUR Iriawan, M.IKom., Ph.D.
Dr. Brodjol Sutijo SU., M.Si.
Irhamah, M.Si, Ph.D

PROGRAM DOKTOR
JURUSAN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2017

LEMBAR PENGESAHAN

Disertasi disusun untuk memenuhi salah satu syarat memperoleh gelar Doktor (Dr.)

di

Program Doktor Departemen Statistika
Institut Teknologi Sepuluh Nopember

Oleh:
Sochardjoepri
NRP. 1310301002

Tanggal Ujian : 22 Juni 2017
Periode Wisuda : September 2017

Disetujui oleh:

1. **Prof. Nur Iriawan, M. Ikom, Ph.D** (Promotor)
   NIP. 19621015 198803 1 002 .....
2. **Dr. Brojol Sutijo S U, M.Si** (Co-Promotor)
   NIP. 19660125 199002 1 001 .....
3. **Irhamah, M.Si, Ph.D** (Co-Promotor)
   NIP. 19780406 200112 2 002 .....
4. **Prof. Ir. Dwi Hendratmo W, M.Sc, Ph.D** (Penguji Eksternal)
   NIP. 19681207 199402 1 001 .....
5. **Dr. Muhammad Mashuri, MT** (Penguji Internal)
   NIP. 19620408 198701 1 001 .....
6. **Dr, Ir, Setiawan, MS** (Penguji Internal)
   NIP. 19601030 198701 1 001 .....
7. **Dr, Bambang W Otok, M.Si** (Penguji Internal)
   NIP. 19681124 199412 1 001 .....

Dekan Fakultas MIPA

Prof. Dr. Basuki Widodo, M.Sc
NIP. 19650605 198903 1 002

IDENTIFIKASI KESAMAAN POLA DOKUMEN TEKS BERDASARKAN KEMUNCULAN *TERM* DALAM KALIMAT

Nama Mahasiswa	: Soehardjoepri
NRP.	: 1310301002
Promotor	: Prof. Drs. NUR Iriawan, M.IKom., Ph.D.
Co-Promotor	: Dr. Brodjol Sutijo SU., M.Si. Irhamah, M.Si., Ph.D.

ABSTRAK

Disertasi ini bertujuan untuk membuat alat deteksi kesamaan pola dokumen teks berdasarkan munculnya *term* di setiap kalimat dalam dokumen teks. Pola munculnya *term* yang diteliti meliputi 3 skenario yaitu: pola munculnya *term* pertama, pola munculnya dua *term* pertama, dan pola munculnya tiga *term* pertama di setiap kalimat dalam dokumen teks. Hasil yang diperoleh berupa cara identifikasi dan kesamaan pola dokumen teks dari munculnya *term* pertama dengan pendekatan uji pembeda pola *Kolmogorov-Smirnov* (uji K-S), dari munculnya dua *term* pertama dengan menghitung jarak *Euclidean* antara pasangan *term* kedua dokumen teks sebagai alat pembeda polanya, dan dari munculnya tiga *term* pertama yang pembedaannya dengan menggunakan pendekatan *Bayesian Network* (BN) dan *likelihood ratio test* dalam dokumen teks.

Pola dokumen teks munculnya *term* pertama dengan pendekatan uji *Kolmogorov-Smirnov* (uji K-S) diperoleh kesamaan pola sebesar 66,67% sesuai skenario dokumen uji. Pola dokumen teks munculnya pasangan *term* pertama dengan menghitung jarak *Euclidean* antara pasangan *term* kedua dokumen teks, diperoleh kesamaan pola sebesar 93,33% sesuai skenario dokumen uji. Sedangkan pola dokumen teks munculnya tiga term pertama dengan pendekatan *Bayesian Network* (BN) dan *likelihood ratio test* dalam dokumen teks diperoleh 100% sama dengan skenario. Ketiga cara pendeteksian pola tersebut terbukti telah mampu membedakan beberapa dokumen standar yang diuji cobakan.

Kata Kunci: pola dokumen teks, munculnya *term, Kolmogorov-Smirnov,* jarak *Euclidean, Bayesian Network, likelihood ratio test*

IDENTIFICATION OF SIMILARITY TEXT DOCUMENTS PATTERN BASED ON TERM APPEARANCE IN SENTENCE

Name	: Soehardjoepri
Registration Number	: 1310301002
Promotor	: Prof. Drs. NUR Iriawan, M.IKom., Ph.D.
Co-Promotor	: Dr. Brodjol Sutijo SU., M.Si.
	Irhamah, M.Si., Ph.D.

ABSTRACT

This dissertation aims to develop a similarity pattern text detection based on the term order appearance in each sentence in the text document. Term emergence patterns examined include three categories, i.e the pattern of the first term emergence, the pattern of the first two terms emergence, and the pattern of the first three terms emergence in each sentence in the text document. The result obtained is the identification and similarity of the text document pattern from the emergence of the first term with the Kolmogorov-Smirnov pattern differentiator approach (KS test), from the appearance of the first two terms by calculating the Euclidean distance between the second term pairs of the text document as a distinguishing tool of the pattern, and from The emergence of the first three terms of distinction by using the Bayesian Network (BN) approach and the likelihood ratio test in text documents.

Pattern of text document the emergence of the first term with Kolmogorov-Smirnov test approach (K-S test) obtained similar pattern of 66.67% according to the test document scenario. The text document pattern of the emergence of the first term pair by calculating the Euclidean distance between the second term pair of text documents, obtained similar pattern of 93.33% according to the test document scenario. While the text document pattern the emergence of the first three terms with the Bayesian Network (BN) approach and the likelihood ratio test in the text document is 100% similar to the scenario.

This dissertation has been succeeded to propose and demonstrate the work of three main algorithms for three scenarios couple with Kolmogorov-Smirnov, Euclidean distance, Bayesian Network and likelihood ratio test respectively to identify and to detect the difference between some standard tested text documents.

Keywords: pattern of text documents, terms of appearance, Kolmogorov-Smirnov, Euclidean distance, Bayesian Network, likelihood ratio test

KATA PENGANTAR

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

Alhamdulillahi Rabbil Alamin, puji syukur dan sujud kehadirat Alloh.SWT, penulis panjatkan kehadirat Alloh.SWT yang telah member ilmu, kesehatan, perlindungan, bimbingan, rahmat, hidayah dan ridho-Nya, sehingga dengan izin-Nya penyusunan disertasi dengan judul "**Identifikasi Kesamaaan Pola Dokumen Teks berdasarkan kemunculan *Term* dalam kalimat**" dapat terselesaikan. Sholawat dan Salam ditujukan kepada Nabi Muhammad.SAW yang telah menuntun umat manusia dari alam gelap gulita ke alam yang terang benderang. Disertasi ini merupakan salah satu persyaratan untuk menyelesaikan pendidikan Program Doktor Jurusan Statistika Institut Teknologi Sepuluh Nopember (ITS) Surabaya.

Penulis mengucapkan terimakasih pada semua pihak atas bimbingan, arahan, bantuan, dorongan moril, bantuan materil, dan do'a hingga terselesaikannya disertasi ini. Pada kesempatan ini penulis menyampaikan penghargaan dan terima kasihkepada:

1. Bapak Rektor ITS yang telah memberikan kesempatan yang diberikan untuk dapat melanjutkan studi ke jenjang pendidikan Program Doktor Statistika di ITS.
2. Bapak Prof. Drs. NUR Iriawan, M.IKom, Ph.D sebagai Promotor yang telah membimbing dan mengarahkan dengan sabar dan tulus dalam menyelesaikan disertasi ini dengan berbagai keterbatasan penulis selama proses penyelesaian disertasi ini. Terimakasih juga atas inspirasi dan pembelajaran hidup yang penulis peroleh selama proses pembimbingan. Semoga Alloh.SWT senantiasa memberikan kesehatan, keberkahan dan rahmat-Nya kepada Bapak Sekeluarga.
3. Bapak Dr. Brodjol Sutijo S U, M.Si dan Ibu Irhamah, M.Si, Ph.D sebagai Co-Promotor atas segala bimbingan, saran, dan diskusi-diskusi yang sangat berharga dalam penyelesaian disertasi ini. Terima kasih atas waktu yang disediakan untuk selalu memberikan dukungan kepada penulis terutama pada

saat-saat yang sangat sulit menjelang penyelesaian disertasi ini. Semoga semua ini menjadi amal ibadah bagi Bapak/Ibu dan semoga Alloh.SWT senantiasa memberikan kesehatan, keberkahan dan rahmat-Nya kepada Bapak/Ibu Sekeluarga.

4. Bapak Dr. Muhammad Mashuri, MT., Bapak Dr. Setiawan, MS., Bapak Dr. Bambang Widjanarko Otok, M.Si,. selaku tim penilai kelayakan yang banyak memberikan saran dan masukan untuk perbaikan disertasi ini.
5. Bapak Bapak Dr. R. Moh. Atok, M.Si dan Bapak Dr. Suhartono, M.Sc selaku tim penilai validasi data yang telah banyak memberikan saran dan masukan dalam perbaikan disertasi ini.
6. Bapak Prof. Ir. Dwi Hendratmo Widyantoro, Ph.D. selaku penguji eksternal yang banyak memberikan saran dan masukan untuk perbaikan disertasi ini.
7. Bapak Ketua Jurusan Statistika, Staf Pengajar, TU dan Karyawan di Jurusan Statistika, di FMIPA, dan di Pascasarjana ITS Surabaya.
8. Bapak Ketua Program Studi S2/S3 Statistika ITS Surabaya.
9. BPP-DN dan DP2M DIKTI yang telah memberikan bantuan beasiswa dan bantuan penelitian.
10. Kepada kedua orang tua (alm), mertua (alm), istri, anak-anak, menantu dan cucu tercinta, serta seluruh keluarga besar di Jember, Jakarta, dan Surabaya, yang telah memberikan motivasi, do'a, dan restu hingga disertasi ini dapat terselesaikan.
11. Segenap rekan-rekan seperjuangan di Program Doktor Jurusan Statistika ITS, terima kasih atas bantuan, kerjasama, kekompakan, persaudaraan, dan kebersamaan selama masa studi.
12. Semua pihak yang telah membantu dalam penulisan disertasi ini yang tidak dapat disebutkan satu persatu.

Pada akhirnya penulis berharap disertasi ini dapat memberikan manfaat dan sumbangan dalam menambah wawasan keilmuan bagi semua pihak. Aamiin ya Rabbal 'alamiin.

Surabaya, Juli 2017

Penulis

DAFTAR ISI

Halaman

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMBANG DAN ARTI

W_{ij}	:	Bobot kata *term* ke-j dan dokumenke-i
f_{ij}	:	Jumlah kemunculan *term* ke-*j* dalam dokumen ke –*i*
N	:	Jumlah semua dokumen yang ada dalam database
k	:	Jumlah dokumen yang mengandung *term* ke-*j*
$\mathbf{A}_{mn}$	:	Matriks yang didekomposisi ukuran *m* x *n*
$\mathbf{U}_{mn}$	:	Matriks ortogonal (matriks vektor singular kiri) ukuran *m* x *n*
$\mathbf{S}_{nn}$	:	Matriks diagonal (matriks nilai singular) ukuran *n* x *n*
$\mathbf{V}_{nn}^{\mathrm{T}}$	:	Transpose matriks orthogonal ukuran *n* x *n*
λ	:	Parameter Distribusi *Poisson*
$\hat{F}(y)$	:	Fungsi Distribusi Komulatif (CDF) empiris
$F(y)$	:	Fungsi Distribusi Komulatif (CDF) hipotesa
$d_{\lvert AB\rvert}$	:	Jarak antara titik *A* ke titik *B*
T_{i}	:	*Term* ke i, i = 1, 2, …, n

DAFTAR LAMPIRAN

BAB 1
PENDAHULUAN

Dalam pendahuluan ini dibahas latar belakang penelitian mengenai identifikasi pola dokumen teks berdasarkan munculnya *term* dalam kalimat di setiap dokumen teks. Penelitian tentang pola dokumen teks akan dibahas berdasarkan munculnya *term* dalam kalimat di setiap dokumen, yaitu dengan memilih 3 skenario dalam deteksi kesamaan pola dokumen teks melalui: munculnya *term* pertama dengan pendekatan uji *Kolmogorov-Smirnov* (uji K-S) dalam dokumen teks, munculnya pasangan *term* pertama dengan menghitung jarak *Euclidean term* dalam dua dokumen teks, dan munculnya tiga *term* pertama dengan pendekatan *Bayesian Network* (BN) dalam dokumen teks. Berdasarkan latar belakang tersebut dirumuskan pokok-pokok permasalahan dalam penelitian ini. Selanjutnya dikemukakan juga tentang tujuan, manfaat, kontribusi, orisinalitas dan batasan penelitian.

1.1 Latar Belakang

Perkembangan teknologi internet dewasa ini sangat pesat. Hal ini diiringi juga dengan semakin berkembangnya teknologi informasi yang dibutuhkan oleh pengguna, sehingga mengakibatkan munculnya suatu cabang ilmu baru dalam teknologi informasi, yaitu pencarian informasi (*information retrieval*). Perkembangan teknologi informasi tentu akan membawa dampak positif dan negatif. Dampak positifnya memudahkan seseorang untuk mencari, melihat dan mempelajari suatu dokumen. Adapun dampak negatifnya adalah membuat seseorang mudah untuk melakukan *copy-paste* sebagian atau keseluruhan dokumen. Jika *copy-paste* dilakukan dengan cara yang benar dengan menyebutkan siapa penulisnya, maka tidaklah menjadi kategori *plagiarisme*, namun sebaliknya jika tidak dilakukan dengan cara yang benar, bisa dikategorikan melakukan *plagiarisme*. *Plagiarisme* merupakan suatu tindakan mengambil sebagian atau keseluruhan karya seseorang dan kemudian mengakuinya sebagai karya sendiri. Hal ini membuat beberapa peneliti untuk menciptakan alat

pendekteksian *plagiarisme* dengan berbagai algoritma, misalkan Algoritma *Smith-Waterman* (Thalib, 2010), Algoritma *Rabin-Karp* (Yoga, 2012), dan Algoritma *Levenshtein Distance* (Hendri, 2012).

Sastroasmoro (2006) mengatakan bahwa klasifikasi atau jenis-jenis *plagiarisme* mencakup: (1). Jenis *plagiarisme* berdasarkan aspek yang dicuri yaitu ide, isi (data penelitian), kata atau paragraph, dan keseluruhan, (2). Klasifikasi berdasarkan *plagiarisme* sengaja atau tidak sengaja, (3). Klasifikasi berdasarkan proporsi atau persentasi kata, kalimat, dan paragraph yang dibajak yaitu ringan: kurang dari 30%, sedang: antara 30 – 70% dan berat atau total: lebih dari 70%, (4). Berdasarkan pada pola *plagiarisme* kata demi kata (*word for word plagiarizing*) dan plagiarisme mosaik. Pendeteksian *plagiarisme* juga dilakukan oleh Ardiansyah (2011) dengan menggunakan metode *Latent Semantic Analysis* (LSA) dan menghitung nilai *similarity* antara dokumen teks satu dengan dokumen teks yang lainnya. Nilai *similarity* didapatkan dari perhitungan *cosine similarity*. Selain itu Kasim (2012) membuat salah satu aplikasi *software* untuk mendeteksi *plagiarisme* dengan menggunakan metode LSA. LSA digunakan untuk menemukan *term-term* yang ada dalam dokumen teks. *Term-term* yang didapatkan dari dokumen teks, dapat digunakan untuk mendeteksi kesamaan dokumen. LSA memiliki kontribusi yang signifikan dalam mendeteksi kesamaan dokumen. Kemampuan untuk mengekstrak dan mewakili makna kontekstual statistik penggunaan kata dalam dokumen teks dapat diterapkan pada corpus besar (Landauer dkk., 1998).

Saat ini banyak sekali dokumen terutama dokumen teks yang bisa dilihat atau diunduh, sehingga dimungkinkan dokumen teks satu dengan dokumen teks yang lainnya mirip atau mempunyai pola yang mirip bahkan bisa terjadi 2 dokumen mempunyai pola yang sama. Untuk mengantisipasi adanya *plagiarisme* dokumen teks, maka perlu adanya sistem/alat yang mampu mendeteksi kesamaan pola dokumen teks. Selain menggunakan hasil *parsing text* dalam kalimat setiap dokumen yang digunakan untuk mendeteksi kesamaan pola munculnya *term* dalam kalimat di setiap dokumen teks, penggunaan uji *Kolmogorov-Smirnov* (uji K-S), jarak *Euclidean*, dan pendekatan *Bayesian Network* (BN), mendesak untuk diteliti sebagai alat pembeda pola dokumen teks.

Uji K-S adalah uji yang digunakan untuk mengetahui apakah suatu data mengikuti pola atau distribusi tertentu. Konsep dasar uji K-S adalah membandingkan fungsi kumulatif pola data atau komulatif distribusi fungsi empiris $\hat{F}(x)$, dengan fungsi distribusi hipotesa $F(x)$ atau kumulatif pola data acuan. Uji K-S akan digunakan untuk menghitung perbandingan antara frekuensi munculnya *term* pertama setiap dokumen teks dengan dokumen teks yang ditentukan sebagai acuan.

Jarak *Euclidean* digunakan untuk menghitung jarak munculnya pasangan *term* dari dua dokumen teks. Masing-masing dokumen mempunyai *term* yang munculnya *term-term* diurutkan dari munculnya *term* ke-1, ke-2 dan seterusnya. Munculnya *term* ke-1 dan ke-2 akan menjadi pasangan *term* di masing-masing dokumen teks. Pasangan *term* dan frekuensi munculnya *term* digambarkan dalam koordinat kartesian 3 dimensi (sumbu *X*, *Y* dan *Z*) yang membentuk titik disuatu permukaan bidang tiga dimensi. Titik-titik pasangan *term* di masing-masing dokumen dapat dihitung jaraknya antara dua dokumen teks dihamparan permukaan bidang tiga dimensi tersebut.

Struktur munculnya *term* (*term-term* yang dihasilkan oleh proses *parsing text* setiap dokumen) dalam setiap kalimat yang mengandung *term-term* tersebut, merupakan order *term* yang tersusun dalam *network*. Munculnya *term* ini didefinisikan sebagai simpul. Jaringan simpul mewakili order munculnya *term* dalam kalimat masing-masing dokumen merupakan peristiwa BN, dikarenakan munculnya *term* yang di depan tergantung munculnya *term* sebelumnya (munculnya *term* yang di depan bersyarat munculnya *term* sebelumnya). Munculnya *term* pertama di setiap kalimat dapat dihitung probabilitasnya di antara semua munculnya *term-term* yang pertama. Munculnya *term* ke dua dapat dihitung probabilitasnya di antara munculnya *term-term* yang ke dua. Probabilitas munculnya *term* berikutnya dapat dihitung probabilitasnya dengan cara yang sama. Probabilitas munculnya *term* yang mewakili order munculnya *term* pertama sampai munculnya *term* terakhir di setiap kalimat dapat dihitung menggunakan probabilitas bersyarat.

Bayesian Network dapat mengandung *n* simpul, yaitu simpul X_1 ke X_n, yang diambil dari kamus *term* (kumpulan *term* yang dihasilkan oleh *parsing text* dalam

dokumen teks standar/uji), maka terjadinya gabungan (*joint*) antar simpul-simpul dapat direpresentasikan sebagai probabilitas $P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$ atau $P(x_1, x_2, \dots, x_n)$. Order munculnya *term* dalam BN menunjukkan order bersyarat munculnya *term* setiap kalimat yang mengandung *term-term* tersebut dalam masing-masing dokumen. Konsep probabilitas bersyarat ini digunakan untuk menghitung probabilitas struktur BN.

Dari latar belakang tersebut, perlu dibuat suatu alat untuk deteksi kesamaan pola dokumen teks berdasarkan munculnya *term* dalam kalimat dengan memilih 3 skenario yaitu: pola munculnya *term* pertama dengan pendekatan uji pembeda pola *Kolmogorov-Smirnov* (uji K-S) dalam dokumen teks, pola munculnya pasangan *term* pertama dengan menghitung jarak *Euclidean term* kedua dokumen teks sebagai alat pembeda polanya, dan pola munculnya tiga *term* pertama yang pembedaannya dengan menggunakan pendekatan *Bayesian Network* (BN) dalam dokumen teks.

1.2 Penelitian Terdahulu

Penelitian terdahulu yang menyangkut plagiarisme telah dilakukan oleh beberapa peneliti diantaranya dengan menggunakan nilai *similarity*, Algoritma *Boyer-Moore*, Algoritma *Knuth-Morris-Pratt*, Algoritma *Rabin Karp*, dan Algoritma *Levenshtein Distance*. Pendekatan sistem deteksi *plagiarisme* yang digolongkan menjadi dua, yaitu *extrinsic* dan *intrinsic.* Pendekatan *extrinsic* bertujuan untuk menemukan kesesuaian bagian teks secara harfiah, sedangkan pendekatan *intrinsic* mencoba untuk mengenali perubahan gaya penulisan (Stein dkk., 2006). Strategi pendekatan *extrinsic* diantaranya perbandingan teks lengkap. Metode ini diterapkan dengan membandingkan semua isi dokumen (Gipp dkk., 2011). Solusi untuk metode ini adalah Algoritma *Boyer-Moore* dan Algoritma *Knuth-Morris-Pratt*. Pendeteksian kesamaan dokumen teks juga ada dalam perancangan sistem deteksi plagiat pada dokumen teks dengan konsep *similarity* menggunakan Algoritma *Rabin Karp* (Salmuasih, 2013).

Winoto (2012) melakukan penelitian bagaimana menerapkan Algoritma *Levenshtein Distance* dalam mendeteksi kemiripan isi dokumen teks. *Levenshtein Distance* merupakan matrik yang digunakan untuk mengukur perbedaan jarak

antara dua urutan kata. *Levenshtein Distance* antara dua *string* ditentukan berdasarkan jumlah minimum perubahan/pengeditan yang diperlukan untuk melakukan transformasi dari satu bentuk *string* ke bentuk *string* yang lain. Algoritma *Levenshtein Distance* yang digunakan dibagi menjadi dua yaitu Algoritma *Levenshtein Distance Standard* dan Algoritma *Lavensthein Distance Preprocessing* (*filtering*, *stemming* dan *sorting*). Winoto (2012) melakukan penelitian ini untuk mengetahui perbandingan kemiripan dokumen yang diperoleh dari *Levenshetin Distance Standard* dan *Lavensthein Distance Preprocessing*. Hasil yang diperoleh Algoritma *Levenshtein Distance Preprocessing* mampu menunjukkan nilai kemiripan tinggi dibandingkan Algoritma *Lavenshtein Distance Standard*.

Nilai *similarity* dari dokumen satu dengan dokumen lainnya mempunyai nilai antara 0 sampai dengan 1. Jika nilai *similarity*-nya semakin mendekati satu maka kedua dokumen tersebut cenderung semakin sama (Ardiansyah, 2011). Pembuatan software untuk pendeteksian *plagiarisme* juga dilakukan oleh beberapa peneliti, diantaranya Kasim (2012) yang membuat aplikasi paket program (*software*) untuk mendeteksi *plagiarisme* dengan menggunakan metode LSA.

Drew (1999) menggunakan algoritma uji K-S untuk menghitung seluruh atau sebagian kumpulan *term* dengan pendekatan *Cumulative Distributin Function* (CDF) dan pengujian K-S, untuk menemukan *p_value* yang terkait dengan uji hipotesis. Uji K-S satu dimensi merupakan metode statistik non-parametrik yang dapat digunakan untuk membandingkan dua distribusi empiris yang mendefinisikan perbedaan mutlak terbesar antara dua fungsi distribusi kumulatif sebagai ukuran perbedaan. Penelitian Drew (1999) menekankan pada perhitungan uji K-S dalam suatu distribusi frekuensi munculnya setiap *term* dalam dokumen. Hal ini memberikan inspirasi peneliti untuk melakukan perhitungan uji K-S pada pola munculnya *term* dalam suatu dokumen teks.

Hendarko (2015) membuat penelitian tentang "Identifikasi citra sidik jari menggunakan alih ragam *wavelet* dan jarak Euclidean" yang bertujuan untuk membuat suatu perangkat lunak yang bisa mengenali pola sidik jari secara automatis dengan menggunakan jarak Euclidean. Kusner dkk. (2015) meneliti

tentang jarak dengan menyajikan Word Mover's Distance (WMD) yang merupakan jarak kata antara dua dokumen teks. WMD mengukur jarak antara dua dokumen teks sebagai jarak minimum dari kata akhir satu dokumen ke kata dalam dokumen lain. Hendarko (2015) menggunakan jarak *Euclidean* untuk dapat mengenali pola sidik jari dan Kusner dkk. (2015) menggunakan jarak *Euclidean* untuk menghitung jarak kata antara dua dokumen, kedua tulisan ini menginspirasi untuk menghitung jarak munculnya *term* antara dua dokumen untuk mengidentifikasi kesamaan pola kemuculan pasangan *term* kedua dokumen.

Model BN yang diusulkan Luis dkk. (2009) difokuskan pada klasifikasi dokumen, menggunakan deskriptor dari tesaurus (kumpulan sinonim atau padan kata). Namun BN, bisa juga digunakan dalam masalah klasifikasi lain di mana kelas yang berbeda telah dikaitkan beberapa jenis deskriptif teks (yang akan memainkan peran deskriptor), misalnya masalah mengklasifikasi dokumen ke dalam direktori web hirarkis. Selain itu, model BN juga bisa digunakan dengan sedikit modifikasi dalam masalah klasifikasi teks hirarkis, asalkan dokumen dapat dikaitkan dengan kategori internal, dengan menghapus node kesetaraan virtual (serta node descriptor). Hal ini menginspirasi membuat pendekatan BN untuk mengidentifikasi kesamaan dokumen dengan pola munculnya tiga term pertama dalam dokumen teks. Semua karya di atas didahului dengan kemampuan memisah *term* dalam kalimat dan menghitung frekuensinya. Pemanfaatan lain dari penghitungan kata-kata dalam dokumen teks tersebut dapat digunakan sebagai bentuk klasifikasi dokumen seperti yang telah dilakukan oleh Grim dkk. (2008) dan Ogura dkk. (2013).

Adapun pendeteksian *plagiarisme* yang menggunakan pola munculnya *term* yang dihasilkan oleh *parsing text* terhadap kalimat di suatu dokumen teks belum pernah diteliti sebelumnya. Untuk itu dibuat alat pendeteksian kesamaan dokumen teks dengan memanfaatkan pola munculnya *term* dalam suatu dokumen teks. Adapun pola dokumen teks berdasarkan munculnya *term* dalam kalimat dengan memilih 3 skenario yaitu: (i) pola munculnya *term* pertama dengan pendekatan uji pembeda pola *Kolmogorov-Smirnov* (uji K-S) dalam dokumen teks, (ii) pola munculnya pasangan *term* pertama dengan menghitung jarak *Euclidean term* kedua dokumen teks sebagai alat pembeda polanya, dan (iii) pola munculnya tiga

term pertama yang pembedaannya dengan menggunakan pendekatan *Bayesian Network* (BN) dalam dokumen teks.

1.3 Perumusan Masalah

Berdasarkan latar belakang penelitian di atas maka dapat ditarik rumusan masalah sebagai berikut :

1. Bagaimana mengidentifikasi dan membedakan pola dokumen teks melalui munculnya *term* pertama dengan pendekatan uji K-S dalam dokumen teks?
2. Bagaimana mengidentifikasi dan membedakan pola dokumen teks melalui munculnya pasangan *term* pertama dengan menghitung jarak *Euclidean term* kedua dokumen teks?
3. Bagaimana mengidentifikasi dan membedakan pola dokumen teks melalui munculnya tiga *term* pertama dengan pendekatan BN dalam dokumen teks?

1.4 Tujuan Penelitian

Capaian penelitian ini ditujukan untuk :

1. Memperoleh cara identifikasi dan cara membedakan pola dokumen teks melalui munculnya *term* pertama dengan pendekatan uji K-S dalam dokumen teks.
2. Memperoleh cara identifikasi dan cara membedakan pola dokumen teks melalui munculnya pasangan *term* pertama dengan menghitung jarak *Euclidean term* kedua dokumen teks.
3. Memperoleh cara identifikasi dan cara membedakan pola dokumen melalui munculnya tiga *term* pertama dengan pendekatan BN dalam dokumen teks.

1.5 Manfaat Penelitian

Keberhasilan pendeteksian pola dokumen teks dalam penelitian ini memberikan manfaat sebagai berikut:

1. Sebagai alternatif pendeteksian kesamaan dokumen teks dengan identifikasi pola dokumen teks berdasarkan munculnya *term-term* dalam kalimat pada dokumen teks.

2. Memberikan kontribusi pada penelitian linguistik dengan pola munculnya *term* sebagai alat identifikasi *plagiarisme*, pola penulisan seseorang dalam dokumen dan pola penulisan dokumen masing-masing daerah.
3. Memberikan gambaran bahwa pendeteksian kesamaan dokumen teks dapat ditentukan melalui frekuensi *term* yang muncul dalam dokumen teks, yang sekaligus menunjukkan pola munculnya *term* dalam kalimat pada dokumen teks.

1.6 Orisinalitas Penelitian

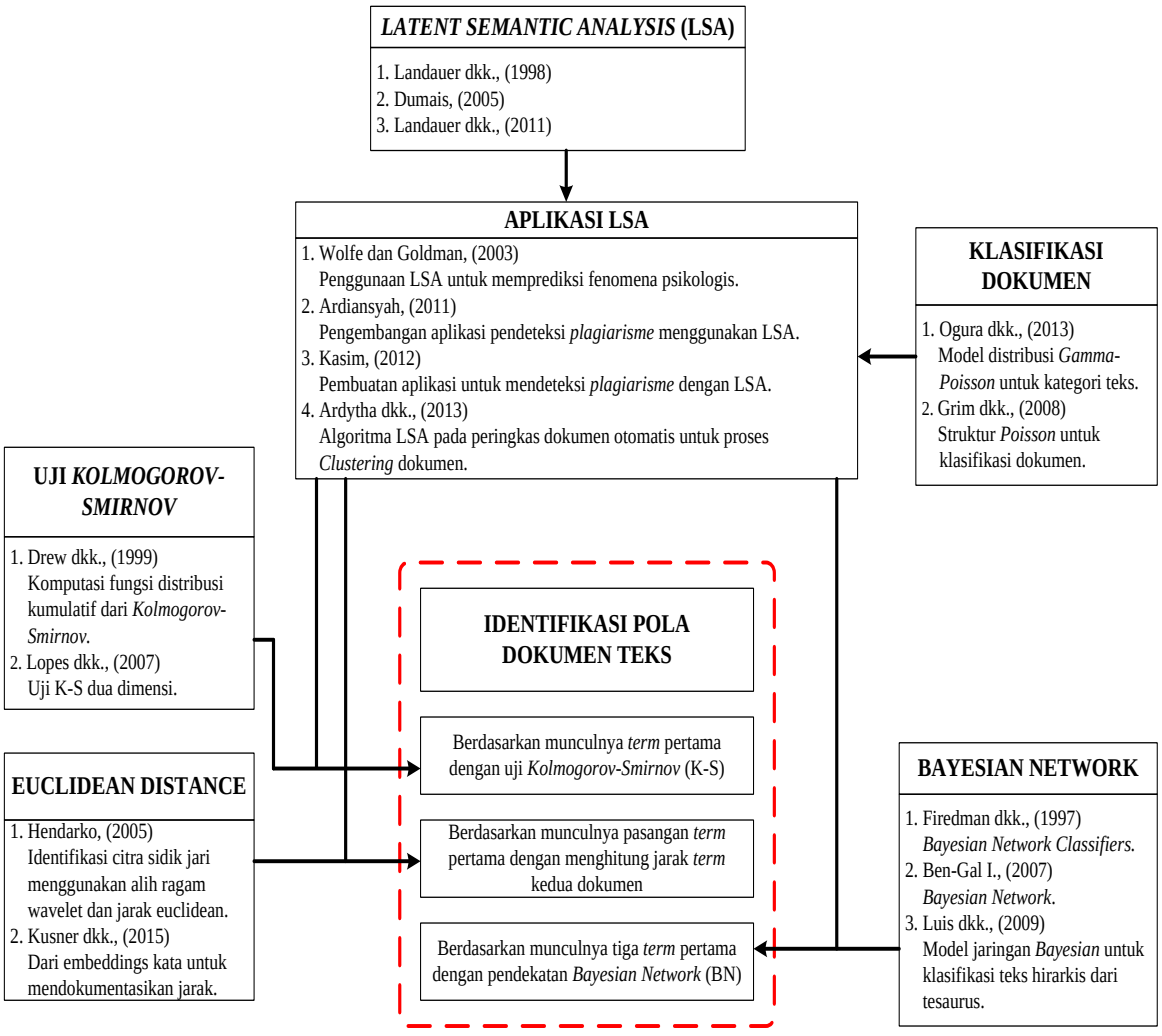
Orisinalitas penelitian yang belum dilakukan oleh peneliti lainnya dapat dilihat dalam Gambar 1.1, dan kontribusi orisinalitas penelitian ini adalah sebagai berikut:

1. Munculnya *term* pertama dalam setiap kalimat terbukti membentuk pola dokumen teks. Dengan menerapkan Algoritma Tabel Munculnya *Term* dan Algoritma Pembuatan Kumulatif Peluang Empiris Munculnya *Term* Pertama, pola dokumen teks tersebut dapat diidentifikasi dan dibedakan antara dokumen satu dengan yang lainnya.
2. Rentetan pasangan *term* pertama dalam setiap kalimat terbukti membentuk pola dokumen teks. Dengan menerapkan Algoritma Kamus *Term*, Algoritma Pasangan *Term*, Algoritma Menghitung Jarak *Term*, dan Algoritma Acuan Kesamaan Dokumen, pola dokumen teks tersebut dapat diidentifikasi dan dibedakan antara dokumen satu dengan yang lainnya.
3. Rentetan tiga *term* pertama dalam setiap kalimat terbukti membentuk pola dokumen teks. Dengan menerapkan Algoritma Order Munculnya *Term*, Algoritma Pola Munculnya *Term*, Algoritma *Likelihood* Munculnya Tiga *Term* Pertama, dan Algoritma Rasio *Likelihood* Dokumen, pola dokumen teks tersebut dapat diidentifikasi dan dibedakan antara dokumen satu dengan yang lainnya.

1.7 Batasan Penelitian

Untuk mempertajam pembahasan dan agar lebih fokus dalam analisis pendeteksian kesamaan pola sebagai tujuan utama penelitian ini, maka diperlukan batasan penelitian sebagai berikut:

1. Kalimat yang dimaksud dalam suatu dokumen merupakan kalimat-kalimat yang signifikan untuk menghasilkan *term-term* oleh proses *parsing text* yang merupakan bagian dari LSA.
2. Dokumen teks yang digunakan Bahasa Indonesia, tidak menguji data berupa gambar maupun suara.
3. Data yang diuji berupa file dokumen format *Microsoft word extension doc* dan *txt* yang telah ditentukan sebagai dokumen standar uji.



Gambar 1.1 Skema orisinalitas penelitian

BAB 2
KAJIAN PUSTAKA DAN DASAR TEORI

Dalam bab ini dibahas kajian teori mengenai beberapa konsep yang dapat dijadikan dasar teori untuk menunjang penelitian ini, yaitu meliputi proses *parsing text* dalam kalimat suatu dokumen yang menghasilkan *term-term*, uji *Kolmogorov-Smirnov* (uji K-S), jarak *Euclidean* antar *term* dari kedua dokumen teks, dan *Bayesian Network* (BN).

2.1 *Parsing Text*

Parsing adalah sebuah proses yang dilakukan seseorang untuk mengupas isi atau maksud suatu kalimat dengan cara memecah kalimat tersebut menjadi kata-kata atau frase-frase. *Parsing text* di dalam pembuatan aplikasi dokumen yang semula berupa kalimat-kalimat berisi kata-kata dan tanda pemisah antar kata seperti titik (“.”), koma (“,”), spasi (“ “) dan tanda pemisah lainnya menjadi kata-kata saja baik itu berupa kata-kata penting maupun tidak penting.

Parsing text dibagi menjadi tiga bagian seperti:

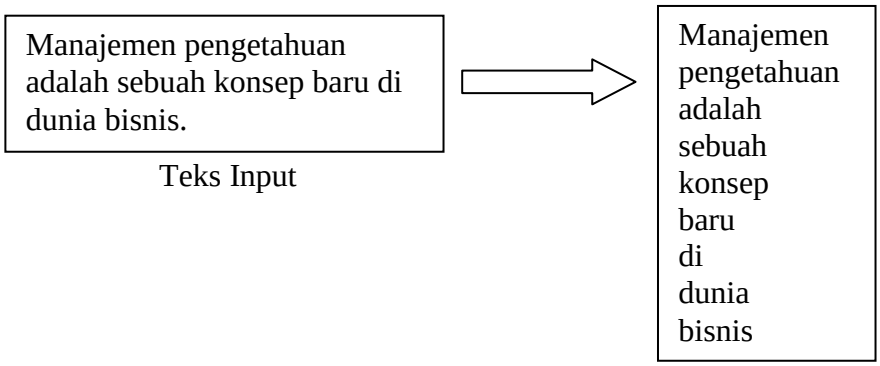
a. *Tokenizing*

Tokenizing merupakan proses mengidentifikasi unit terkecil (*token*) dari suatu struktur kalimat. Tujuan dilakukannya *tokenizing* ini adalah untuk mendapatkan *term-term* (kata unik) yang nantinya akan diindeks. Pengklasifikasian token dilakukan untuk teks yang dipisahkan dengan spasi atau atau “*carriage return*” dalam suatu dokumen. Adapun beberapa kasus yang ditangani oleh *tokenizing* yaitu:

1) *Handling Special Character*, untuk pengambilan polanya menggunakan *regular expression*.
2) *Phrase* merupakan kelompok kata yang saling berkaitan namun tidak mengandung unsur subject dan verb. Dengan memahami bagaimana cara membentuk dan fungsinya, akan memudahkan seorang penulis untuk membuat variasi di dalam suatu tulisan. Selain spesial karakter,

tokenizing juga dapat menangani beberapa pola *phrase* seperti nama, tempat dan kata sifat.

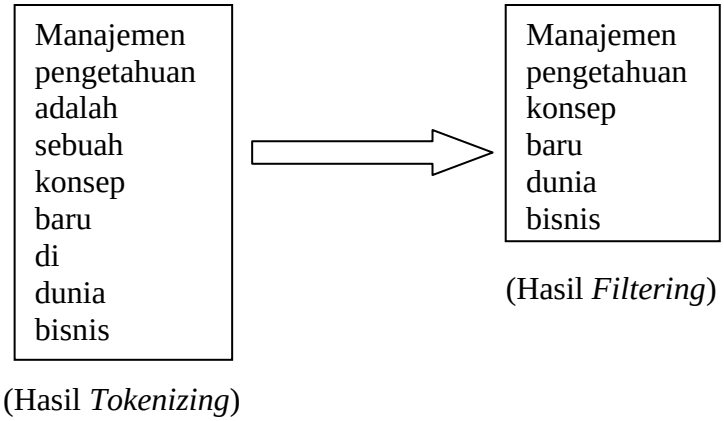
3) *Whitespace*, karakter ini diabaikan oleh *tokenizing* dan dianggap sebagai pemisah token. Gambar 2.1 adalah proses *Tokenizing*. (Triawati, 2009)



Gambar 2.1 *Tokenizing*

b. Filtering

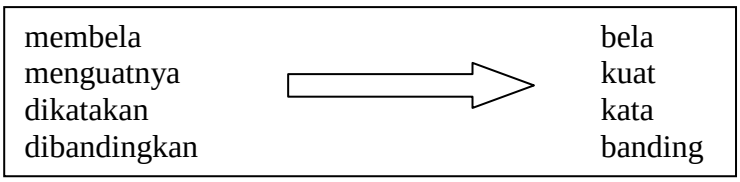
Filtering adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting). *Stoplist*/*stopword* adalah kata umum (*common words*) yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna. *Stopwords* umumnya dimanfaatkan dalam *task information retrieval.* Contoh *stopwords* adalah "yang", "dan", "di", "dari", seterusnya. Contoh dari tahapan ini dapat dilihat pada Gambar 2.2: (Triawati, 2009)



Gambar 2.2 *Filtering*

c. Stemming

Stemming merupakan suatu proses untuk menemukan kata dasar dari sebuah kata. Dengan menghilangkan semua imbuhan (*affixes*) baik yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan *confixes* (kombinasi dari awalan dan akhiran) pada kata turunan. *Stemming* digunakan untuk mengganti bentuk dari suatu kata menjadi kata dasar dari kata tersebut yang sesuai dengan struktur *morfologi* Bahasa Indonesia yang baik dan benar. Contoh dari tahapan ini pada teks adalah seperti Gambar 2.3: (Triawati, 2009)

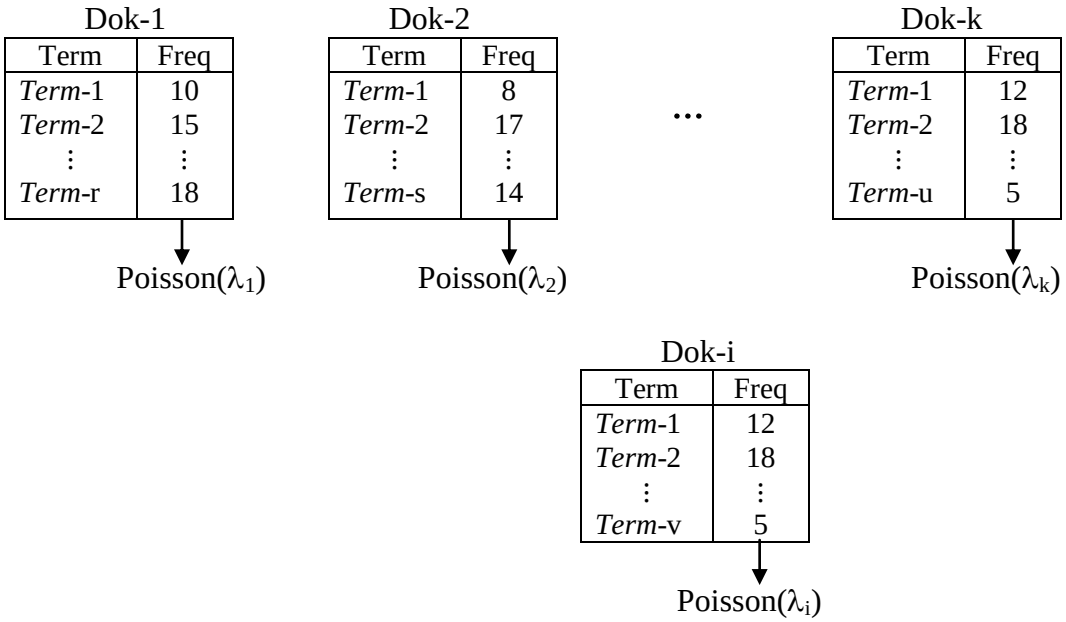


(Hasil *Filtering*) (Hasil *Stemming*)

Gambar 2.3 *Stemming*

2.2 Frekuensi Munculnya *Term* Dalam Dokumen

Munculnya *term* di setiap kalimat dalam dokumen teks mempunyai arti. Frekuensi munculnya *term* yang ada dalam dokumen teks akan dapat membedakan pola antara dokumen satu dengan yang lainnya. Ogura dkk. (2013) memperkenalkan model baru untuk menggambarkan distribusi frekuensi *term* dalam dokumen untuk klasifikasi teks dengan menggunakan distribusi *Gamma-Poisson* untuk merepresentasikan frekuensi *term* sebagai model teks yang lebih baik. Ilustrasi model distribusi *Gamma-Poisson* untuk kategori teks dapat dilihat dalam Gambar 2.4, dengan rincian sebagai berikut: misalkan ada sejumlah dokumen (Dok-1, Dok-2, …, Dok-i, …, Dok-k) yang masing-masing dokumen mempunyai frekuensi munculnya *term* (*term*-1, *term*-2, …, *term*-T). Frekuensi *term* dalam setiap dokumen akan berpola *Poisson*(λ_i), i = 1, 2, …, *k*. Rentetan λ_i, i = 1, 2, …, *k* mengikuti pola/distribusi *Gamma*(α,β) atau $\lambda_i \sim$ *Gamma*(α,β).



Gambar 2.4 Ilustrasi model distribusi *Gamma-Poisson*

2.3 Uji *Kolmogorov-Smirnov*

Uji *Kolmogorov-Smirnov* (uji K-S) adalah uji yang digunakan untuk mengetahui apakah suatu data mengikuti pola atau distribusi tertentu. Konsep dasar uji K-S adalah membandingkan fungsi distribusi kumulatif (CDF) empiris $\hat{F}(x)$, dengan fungsi distribusi kumulatif hipotesa $F(x)$.

Apabila terdapat urutan *term* pertama yang dapat dipisahkan dari kalimat-kalimat (kalimat yang mengandung *term-term* hasil *parsing text*) dalam dokumen sebagai suatu order sampel $X_1, X_2, \dots, X_q$ yang dapat disusun sebagai bentuk pola CDF empiris $\hat{F}_A(x)$ dengan

$$\hat{F}_A(x) = \begin{cases} 0, & x < x_1 \\ \dfrac{b}{q}, & x_b \le x < x_{b+1}, \quad b = 1, 2, \dots, q-1 \\ 1, & x \ge x_q \end{cases} \tag{2.3}$$

dan jika order sampel tersebut diduga mempunyai pola yang sama dengan order sampel *term* pertama dari kalimat-kalimat dokumen lain namakan $F_B(x)$, maka uji K-S dapat digunakan untuk menguji kesamaan pola kedua dokumen tersebut. Selanjutnya statistik uji yang digunakan adalah jarak vertikal tersebar

(maksimum) antara $\hat{F}_A(x)$ dan $F_B(x)$ yang dinotasikan dengan D_q, sehingga (Lehman dan Romano, 2005):

$$D_q = \max\left|\hat{F}_A(x) - F_B(x)\right|, \tag{2.4}$$

dengan hipotesis yang digunakan adalah:

H_0: $\hat{F}_A(x) = F_B(x)$

H_1: $\hat{F}_A(x) \neq F_B(x)$.

Kriteria pengambilan keputusan K-S, H_0 akan ditolak jika

$$\left(\sqrt{q} + 0{,}12 + \frac{0{,}11}{\sqrt{q}}\right) D_q > c_{1-\alpha}$$

dengan $c_{1-\alpha}$ = 1,224 untuk $\alpha = 10\%$ yang terdapat dalam Tabel 2.1 (Law, 2000). Dari uraian diatas dapat dibuat Algoritma Perhitungan D_q dengan rincian seperti dalam Algoritma 2.1.

Tabel 2.1. Modifikasi nilai kritis $c_{1-\alpha}$ untuk uji statistik K-S (Law, 2000)

Kasus	Uji statistic K-S	$1-\alpha$				
		0,850	0,900	0,950	0,975	0,990
Semua parameter	$\left(\sqrt{q} + 0{,}12 + \frac{0{,}11}{\sqrt{q}}\right) D_q$	1,138	1,224	1,358	1,480	1,628

Algoritma 2.1: Perhitungan D_q

Langkah 1. Ambil q observasi $X_1, X_2, \ldots, X_q$,
Langkah 2. Menghitung nilai $\hat{F}_A(x)$ dan $F_B(x)$,
Langkah 3. Membandingkan nilai $\hat{F}_A(x)$ dan $F_B(x)$,
Langkah 4. Menghitung $D_q = \max\left|\hat{F}_A(x) - F_B(x)\right|$.

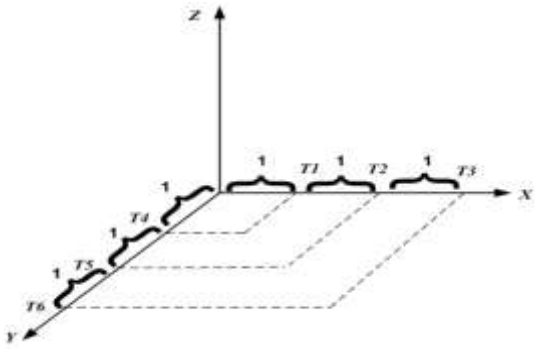
2.4 Jarak *Euclidean*

Jarak *Euclidean* antara dua buah titik adalah panjang garis yang menghubungkan kedua titik itu. Ilustrasi jarak *Euclidean*, ambil titik $A(x_A, y_A, z_A)$ di dokumen kesatu, dengan sumbu *absis* x_A: munculnya *term* pertama, sumbu *ordinat* y_A: munculnya *term* kedua, dan z_A: besar frekuensi munculnya pasangan *term* pertama dan kedua. Ambil titik $P(x_P, y_P, z_P)$ di dokumen kedua, dengan sumbu *absis* x_P: munculnya *term* pertama, sumbu *ordinat* y_P: munculnya *term*

kedua, dan z_P: besar frekuensi munculnya pasangan *term* pertama dan kedua. Dari kedua titik tersebut dapat dihitung jarak *Euclidean* antara dua buah titik $A(x_A, y_A, z_A)$ dan $P(x_P, y_P, z_P)$ dengan cara sebaga berikut:

$$|AP| = \sqrt{(x_A - x_P)^2 + (y_A - y_P)^2 + (z_A - z_P)^2} \quad (2.5)$$

Skala munculnya *term* digunakan satu-satuan jarak di setiap sumbu. Setiap munculnya *term* pertama setiap kalimat dalam dokumen diletakkan di sumbu X dan di sumbu Y untuk munculnya *term* kedua. Untuk munculnya *term* yang sama, diletakkan pada posisi yang sama. Skala munculnya *term* ini dapat ditunjukkan pada Gambar 2.5.



Gambar 2.5 Skala munculnya *term* di sumbu X dan sumbu Y

Dari uraian tentang jarak *Euclidean* diatas dapat dibuat Algoritma Jarak *Euclidean* dengan rincian seperti dalam Algoritma 2.2.

Algoritma 2.2: Jarak *Euclidean*

Langkah 1.	Ambil *term* $T_i \in$ S, i = 1, … n, S = kumpulan *term* hasil *parsing text*,
Langkah 2.	Menyusun *term-term* {T_1, T_2, …, T_n} dari setiap kalimat di masing-masing dokumen,
Langkah 3.	Membuat tabel munculnya *term* pertama dan kedua,
Langkah 4.	Gambarkan setiap munculnya *term* pertama di sumbu X dan munculnya *term* kedua di sumbu Y dengan skala satu-satuan langkah dari pusat (0,0) dan diikuti satu-satuan langkah berikutnya untuk *term* selanjutnya,
Langkah 5.	Gambarkan tingginya frekuensi pasangan *term* pertama dan kedua di sumbu Z dengan langkah satu satuan keatas,
Langkah 6.	Mengitung jarak pasangan *term* (munculnya *term* pertama dan *term* kedua) dari kedua dokumen teks, dengan menggunakan persamaan (2.5),
Langkah 7.	Lakukan sampai seluruh munculnya pasangan *term* di masing-masing dokumen.

Sebagai contoh jarak *Euclidean*, misal Dok-A dan Dok-B yang berisi data munculnya *term* pertama dan kedua di setiap kalimat untuk kedua dokumen seperti diberikan dalam Tabel 2.2.

Tabel 2.2. Munculnya term pertama dan kedua pada Dok-A dan Dok-B

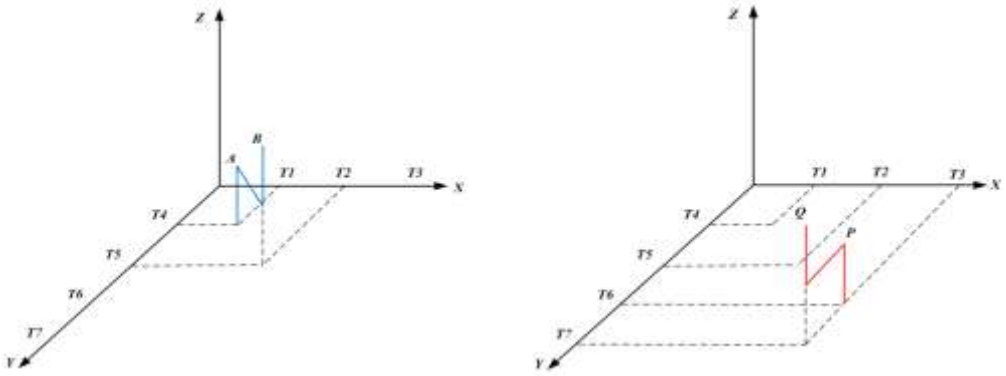
Dok-A	Munculnya *Term* ke-	
Kalimat	1	2
1	T_1	T_4
2	T_2	T_5
3	T_3	T_6
4	T_3	T_7

Dok-B	Munculnya *Term* ke-	
Kalimat	1	2
1	T_3	T_6
2	T_3	T_7
3	T_1	T_4
4	T_2	T_5

Penggambaran titik-titik dalam koordinat Kartesius dengan sumbu *X* untuk munculnya *term* pertama (Dok-A: T_1, T_2, dan T_3, Dok-B: T_3, T_1, dan T_2), sumbu *Y* untuk munculnya *term* kedua (Dok-A: T_4, T_5, T_6 dan T_7, Dok-B: T_6, T_7, T_4 dan T_5) dan sumbu *Z* frekuensi munculnya pasangan *term* pertama dan kedua dalam setiap dokumen, dengan skala satu satuan (satu langkah) setiap munculnya *term* di masing-masing kalimat dalam dokumen dalam sumbu-sumbu koordinat. Misal ambil titik *A*(1,1,1) di dokumen kesatu, dibangun oleh x_A = munculnya *term* T_1 berada pada sumbu *X*, 1 langkah dari pusat (0,0), y_A = munculnya *term* T_4 berada pada sumbu *Y*, 1 langkah dari pusat (0,0), dan z_A = frekuensi munculnya pasangan *term* T_1 dan T_4, sebesar 1 langkah berada pada sumbu *Z*. Misal ambil titik *P*(3,3,1) di dokumen kedua, yang dibangun oleh x_P = munculnya *term* T_3, berada pada sumbu *X*, 3 langkah dari pusat, y_P = munculnya *term* T_6, berada pada sumbu *Y*, 3 langkah dari pusat, dan z_P = frekuensi munculnya pasangan *term* T_3 dan T_6 sebesar 1 langkah berada pada sumbu *Z*. Misal ambil juga *B*(2,2,2) di dokumen kesatu dan *Q*(3,4,2) di dokumen kedua, maka jarak

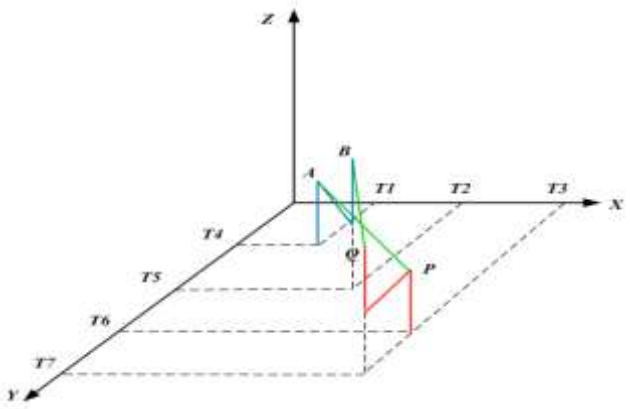
$$d_{|AP|} = \sqrt{(1-3)^2 + (1-3)^2 + (1-1)^2} = 2{,}83 \text{ dan}$$
$$d_{|BQ|} = \sqrt{(2-3)^2 + (2-4)^2 + (2-2)^2} = 2{,}24.$$

Ilustrasi munculnya *term* pertama dan kedua untuk Dok-A dan Dok-B terdapat dalam Gambar 2.6.(a) dan (b), sedangkan jarak *Euclidean* $d_{|AP|}$ dan $d_{|BQ|}$ terdapat dalam Gambar 2.7.



(a). Dok-A (b). Dok-B

Gambar 2.6 Kemunculan *term* pertama dan kedua di Dok-A dan Dok-B



Gambar 2.7 Jarak $d_{|AP|}$ = 2,83 dan $d_{|BQ|}$ = 2,24

2.5 Uji Hipotesis Dengan Membandingkan Fungsi *Likelihood* (*Likelihood Rasio Test*)

Suatu bentuk yang sangat popular dari pengujian hipotesis adalah uji dengan membandingkan fungsi likelihood, yang merupakan generalisasi dari tes optimal untuk hipotesis sederhana yang dikembangkan oleh Neyman dan Pearson. Uji *Likelihood Rasio* (LR) didasarkan pada fungsi kemungkinan $f_n(x_1, \dots, x_n|\theta)$ dan intuisi bahwa fungsi kemungkinan cenderung tertinggi dekat nilai sebenarnya dari θ.

Definisi: (Somayasa, 2008)

Misalkan $L(\theta;\ x_1, \dots, x_n)$ merupakan fungsi *likelihood* dengan variabel random $X_1, \dots, X_n$. Misalkan

$$\Lambda(x_1, \dots, x_n) = \frac{sup_{\theta \epsilon H_0} L(\theta;\ x_1, \dots, x_n)}{sup_{\theta \epsilon \Theta} L(\theta;\ x_1, \dots, x_n)}$$

Tes LR berukuran α untuk hipotesis $H_0: \theta \epsilon \Theta_0$ vs $H_1: \theta \epsilon \Theta_1$adalah`

$$\Phi(x_1, \dots, x_n) = \begin{cases} 1; & jika\ \Lambda(x_1, \dots, x_n) \leq k \\ 0; & jika\ \Lambda(x_1, \dots, x_n) > k \end{cases}$$

dimana $0 < k < 1$ adalah konstanta yang tidak diketahui yang ditentukan dari persamaan

$$sup_{\theta \epsilon H_0}\ \mathbb{P}\{\Lambda(x_1, \dots, x_n) \leq k\} = \alpha.$$

2.6 Metode *Bayesian*

Metode *Bayesian* diadopsi dari nama penemu metode tersebut yaitu Thomas Bayes (1702-1761). Namun demikian, metode *Bayesian* baru mulai dikenal pada tahun 1764 setelah Thomas Bayes meninggal dunia. Walaupun metode *Bayesian* sudah ada sejak abad 18, namun sampai dengan awal abad 20, metode *Bayesian* kurang populer dibandingkan dengan metode klasik (*frequentist*). Penggunaan metode *Bayesian* mulai meningkat seiring perkembangan teknologi informasi dan semakin luas penggunaan komputer sehingga analisis data yang sulit dilakukan secara analitik dapat diperoleh solusinya dengan menggunakan simulasi komputer (Gelman dkk., 2004).

Inferensia dengan pendekatan *Bayesian* berbeda dengan pendekatan klasik meskipun kedua metode tersebut menggunakan fungsi *Likelihood* data. Dalam pendekatan klasik, parameter model yang akan diestimasi diasumsikan bernilai tunggal. Proses estimasi dengan pendekatan klasik, umumnya dilakukan untuk mendapatkan nilai parameter yang dapat memaksimumkan fungsi *Likelihood* yang dianggap sebagai fungsi dari parameter tersebut. Pada kasus yang kompleks, proses estimasi dengan metode klasik tersebut umumnya menggunakan teknik optimasi secara numerik untuk mendapatkan

solusinya. Sementara dalam pendekatan *Bayesian*, seluruh parameter yang tidak diketahui dipandang sebagai variabel random yang dikarakteristikkan oleh distribusi *prior* dari parameter tersebut (Ntzoufras, 2009; Gelman dkk., 2004; Congdon, 2006). Distribusi *prior* dari parameter menyatakan variasi parameter tersebut.

Berbeda dengan pendekatan klasik, metode *Bayesian* tidak melibatkan proses optimasi dalam inferensia, karena pendekatan *Bayesian* mengaplikasikan Teorema *Bayes* yang didasarkan pada distribusi *posterior* gabungan dari seluruh parameter (King, Morgan, Gimenez dan Brooks, 2010). Selanjutnya metode *Bayesian* akan melakukan estimasi parameter dengan menggunakan distribusi *posterior marginal* parameter tersebut. Distribusi *posterior marginal* ini diperoleh dengan cara mengintegralkan distribusi *posterior* gabungan. Pada tahap ini, khususnya untuk kasus model yang cukup komplek umumnya timbul permasalahan yaitu proses integrasi tersebut menjadi sangat rumit dan sulit untuk memperoleh solusinya. Namun, metode *Bayesian* dapat mengatasi permasalahan tersebut. Dalam hal ini, cara yang digunakan dalam pendekatan modern dari analisis *Bayesian* adalah bukan dengan cara mengintegralkan distribusi *posterior* gabungan secara analitik, melainkan dengan menggunakan prosedur simulasi data yang mengikuti distribusi *posterior* gabungan dengan memanfaatkan bentuk *full conditional* untuk memperoleh distribusi *posterior* marginal setiap parameter yang akan diestimasi tersebut. Dengan demikian, proses optimasi yang dilakukan dalam analisis klasik digantikan dengan proses integrasi dalam pendekatan analisis *Bayesian* (King dkk., 2010). Integrasi dalam analisis *Bayesian* ini tidak dilakukan secara analitik terhadap distribusi *posterior* gabungan dari parameter, melainkan dengan pendekatan prosedur simulasi khusus yang menghasilkan sampel dari distribusi *posterior* tersebut. Proses ini selanjutnya dikenal sebagai proses *Markov Chain Monte Carlo* (MCMC) (King dkk., 2010).

Metode *Bayesian* merupakan salah satu metode alternatif untuk mengestimasi parameter model. Ketersediaan paket program untuk analisis Bayesian membuat metode ini menjadi lebih berdayaguna dan fleksibel dalam

analisis pemodelan secara stokastik yang kompleks. Akibatnya, beberapa keterbatasan dalam pemodelan secara klasik dapat diatasi seperti model yang kompleks, asumsi-asumsi yang tidak sesuai dengan realita, dan simplifikasi dapat dihindari (King dkk., 2010).

Perspektif *Bayesian* menyatakan bahwa data hasil pengamatan berasal dari suatu distribusi probabilitas yang memiliki parameter-parameter yang tidak diketahui dengan pasti. Oleh karena itu perlu ditentukan suatu distribusi dari parameter tersebut yang disebut sebagai distribusi *prior*. Kombinasi antara *prior* dari parameter dengan informasi data sampel akan menghasilkan distribusi *posterior* dari parameter. Distribusi *posterior* ini menyatakan pola ketidakpastian dari nilai parameter populasi setelah diperoleh data hasil pengamatan. Pada umumnya *varians* dari distribusi *posterior* ini lebih kecil dibandingkan dengan *varians* dari distribusi *prior* (Gelman dkk., 2004). Hal tersebut menunjukkan bahwa data hasil pengamatan menurunkan tingkat ketidakpastian dari nilai parameter populasi.

Inferensia dengan pendekatan *Bayesian* dilakukan menggunakan distribusi *posterior* dari parameter. Oleh karena itu, tujuan utama untuk metode *Bayesian* adalah melakukan eksplorasi terhadap distribusi *posterior* tersebut. Dalam implementasi, metode *Bayesian* banyak digunakan untuk analisis model statistik yang kompleks (Carlin dan Chib, 1995).

Teorema Bayes

Jika Y adalah variabel random yang mengikuti pola distribusi tertentu dengan fungsi kepadatan peluang (pdf) $f(\boldsymbol{y}|\boldsymbol{\theta})$ dengan $\boldsymbol{\theta}$ adalah vektor parameter berukuran d atau $\boldsymbol{\theta} = [\theta_1 \ \theta_2 \ \cdots \ \theta_d]^T$ dan $\boldsymbol{y} = [y_1 \ y_2 \ \dots \ y_n]^T$ adalah sampel berukuran n yang berdistribusi identik dan independen, maka berdasarkan aturan probabilitas, distribusi gabungan dari $\boldsymbol{\theta}$ dan $\boldsymbol{y}$ dapat diformulasikan dalam bentuk persamaan (2.6): (Box dan Tiao, 1973; Gelman dkk., 2004).

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\boldsymbol{y}) f(\boldsymbol{y}) \tag{2.6}$$

Hal terpenting dalam analisis Bayesian terletak pada penentuan distribusi

posterior $f(\boldsymbol{\theta}|\boldsymbol{y})$. Berdasarkan aturan probabilitas dalam teorema Bayes, distribusi *posterior* dari parameter $\boldsymbol{\theta}$ dapat dinyatakan dalam persamaan berikut:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{y})}, \tag{2.7}$$

dengan $f(\boldsymbol{y}) = \begin{cases} \int \cdots \int f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\theta_1 \cdots d\theta_d & \text{jika } \boldsymbol{\theta} \text{ kontinu} \\ \sum \cdots \sum f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) & \text{jika } \boldsymbol{\theta} \text{ diskrit} \end{cases}$

$f(\boldsymbol{\theta})$ adalah fungsi distribusi prior dari parameter $\boldsymbol{\theta}$,

$f(\boldsymbol{y}|\boldsymbol{\theta})$ adalah fungsi likelihood data yang berisi informasi sampel dan

$f(\boldsymbol{y})$ adalah fungsi densitas.

Sehingga, persamaan (2.7) dapat dinyatakan dalam bentuk proporsional sebagai berikut:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \tag{2.8}$$

Persamaan (2.8) menunjukkan bahwa *posterior* dari parameter $\boldsymbol{\theta}$ diperoleh dari proses pembaruan informasi *prior* $\boldsymbol{\theta}$ dengan menggunakan informasi data hasil observasi melalui fungsi *Likelihood*. Dengan demikian fungsi *Likelihood* data memegang peranan penting dalam aturan Bayes. Fungsi *Likelihood* dapat dinyatakan dengan persamaan (2.9):

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}) \tag{2.9}$$

2.7 *Bayesian Network* (BN)

Menurut Pearl (1988), BN merupakan model *graph* dari sebaran peluang. BN dapat berupa model *spatial-temporal* dimana terdapat ketergantungan dan ketidak-tergantungan antara satu *node* dengan *node* lainnya dalam bentuk *graph* (*Directed Acyclic Graph*/DAG*)* yang mudah dimengerti dan diinterpretasikan (Cofino dkk., 2002). Bahkan, proses perhitungannya sangat efisien dan dapat memberikan informasi peluang dengan sederhana dan lengkap serta merupakan teknik yang populer dalam menyelesaikan suatu masalah jika ketidakpastiannya

sangat kompleks (Mittal dkk., 2007). Menurut Cooper dan Herskovits (1992); Cano dkk. (2004), berdasarkan komponennya, BN terdiri atas *Bayesian Structure* (*structure learning*) dan *Bayesian Parameter* (*parametric learning*).

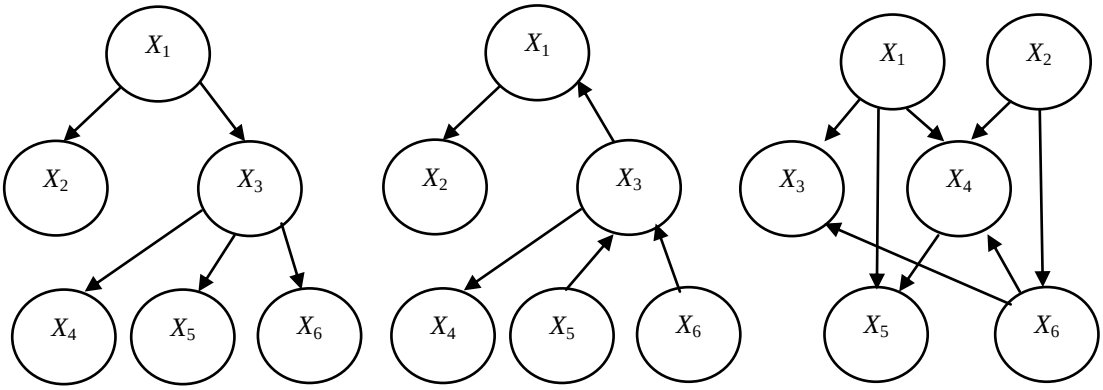
Sementara itu, menurut Cofino dkk. (2002) dan Cano dkk. (2004), komponen dari BN sebagai penilaian dan pencarian yang meliputi:

1. ukuran kualitas, untuk menentukan kualitas struktur grafik dan pendugaan parameter bagi kandidat BN, dan
2. algoritma pencarian, untuk pencarian secara efisien ruang BN yang mungkin dan menemukan satu kualitas yang terbaik.

BN merupakan kepadatan probabilitas gabungan (*joint probability density*) lebih dari satu himpunan variabel $\mathbf{X}$ (Jensen, 1996). Kepadatan gabungan (*joint density*) ditentukan melalui *Directed Acyclic Graph* (DAG). *Directed graph* terdiri dari satu himpunan *node* dan himpunan *arc*. *Arc* (u;v) berawal dari *node* u (induk) ke *node* v (anak). *Path* merupakan penghubung antar *node* yang setiap pasangan *node* berturut-turut berdekatan. *Path* adalah lingkaran yang terdiri dari lebih dari dua *node* dimana *node* pertama dan *node* terakhirnya sama. Lingkaran adalah *directed* jika bisa mencapai *node* yang sama saat mengikuti busur yang berada di arah yang sama. Directed graph adalah *acyclic* (DAG) jika tidak mengandung *directed cycles*. Sebuah *graph* terhubung jika ada *path* antara setiap pasangan *node*. Sebuah *graph* terhubung tunggal jika ada tepat satu *path* antara setiap sepasang node; jika tidak, *graph* adalah *multiply-connected*. Sebuah *graph* terhubung tunggal juga disebut *polytree*. Dalam BN, setiap *node* dari *graph* merupakan variabel acak X_i sebagai anggota dalam $\mathbf{X}$. Induk dari X_i dilambangkan oleh pa(X_i). Semantik dari model BN ditentukan oleh kondisi *Markov*: setiap variabel independen dari *nondescendants* noninduk yang diberikan merupakan induknya. Kondisi ini akan membangun kepadatan probabilitas gabungan yang tunggal (*unique joint probability density*) (Pearl, 1988):

$$p(\mathrm{X}) = \prod_i^n p(X_i | pa(X_i)). \qquad (2.10)$$

Setiap variabel acak X_i dapat dihitung kepadatan probabilitas bersyaratnya yaitu $p(X_i|pa(X_i))$. Ilustrasi DAG sebagai BN dapat dilihat dalam Gambar 2.8.(a). BN sebagai *Tree*, (b). BN sebagai *Polytree* dan (c). BN sebagai *Multi-connected Bayes net*, sebagai berikut:



(a) (b) (c)

Gambar 2.8 BN: (a) *Tree*, (b) *Polytree*, (c) *Multi-connected Bayes net*

Dari uraian tentang BN diatas dapat dibuat Algoritma Perhitungan BN dengan rincian seperti dalam Algoritma 2.3.

Algoritma 2.3: Perhitungan *Bayesian Network*

Langkah 1.	Ambil *term* $T_i \in$ S, i = 1, … n, S = kumpulan *term* hasil *parsing text*,
Langkah 2.	Tandai munculnya *term* hasil *parsing text*,
Langkah 3.	Membuat order munculnya *term* $\{T_1, T_2, \ldots, T_n\}$ di setiap kalimat dari masing-masing dokumen,
Langkah 4.	Ambil munculnya tiga *term* pertama $\{T_1, T_2, T_3\}$ pada order *term* $\{T_1, T_2, \ldots, T_n\}$,
Langkah 5.	Menghitung likelihood munculnya tiga *term* pertama untuk setiap kalimat $P(T_1, T_2, T_3) = P(T_3\|T_2,T_1).P(T_2\|T_1).P(T_1)$,
Langkah 6.	Menghitung probabilitas munculnya tiga *term* pertama dengan pendekatan *Bayesian Network* untuk semua dokumen.

BAB 3
METODE PENELITIAN

Dalam bab ini dibahas dua tahapan penelitian yaitu tahapan kajian teori dan tahapan implementasinya. Kajian teori dilakukan berkaitan dengan *parsing text* dalam kalimat suatu dokumen, Uji *Kolmogorov-Smirnov* (uji K-S), jarak *term* antara dua dokumen dan *Bayesian Network* (BN). Setelah tahapan kajian teori selesai dilanjutkan dengan tahapan implementasi teori tersebut untuk mengidentifikasi pola dokumen teks.

Tahapan Kajian Teori dan Tahapan Implementasi

Pada tahapan kajian teori ada tiga tahapan yang merupakan tujuan dari penelitian ini. Tahapan pertama adalah membuat algoritma-algoritma untuk mengidentifikasi kesamaan pola dokumen melalui munculnya *term* pertama di setiap kalimat (kalimat yang mengandung *term-term* hasil dari proses *parsing text*) dalam dokumen dengan pendekatan uji *Kolmogorov-Smirnov* (uji K-S), tahapan kedua adalah membuat algoritma-algoritma untuk mengidentifikasi kesamaan pola dokumen melalui munculnya pasangan *term* pertama di setiap kalimat dalam dokumen dengan pendekatan jarak *Euclidean* dan tahapan ketiga adalah membuat algoritma-algoritma dan teorema untuk mengidentifikasi kesamaan pola dokumen melalui munculnya tiga *term* pertama di setiap kalimat dalam dokumen dengan pendekatan *Bayesian Network* (BN).

3.1 Tahapan Pola Munculnya *Term* Pertama

Tahapan pola munculnya *term* pertama adalah langkah-langkah untuk menjawab tujuan kesatu yaitu membuat algoritma-algoritma untuk mengidentifikasi kesamaan pola munculnya *term* pertama di setiap kalimat dalam dokumen teks dengan pendekatan Uji K-S, dengan rincian tahapan sebagai berikut:

(a) Melakukan proses *parsing text* pada 6 dokumen standar yaitu Dok-1, Dok-2, Dok-3, Dok-4, Dok-5 dan Dok-6, merupakan dokumen teks yang diuji, untuk menghasilkan *term-term* pada setiap dokumen teks.

(b) Menghitung frekuensi munculnya *term* untuk masing-masing dokumen.

(c) Membuat kode *term* untuk masing-masing dokumen.

(d) Membuat Algoritma Kamus *Term*, dengan rincian tahapan sebagai berikut:

- Siapkan sebanyak K dokumen akan diuji kesamaannya.
- Namakan masing-masing dokumen terorder dari dokumen 1 sampai dengan dokumen K.
- Proseslah semua K dokumen dengan proses *parsing text* menghasilkan *term-term* dari kalimat yang berada di masing-masing dokumen.
- Membuat daftar order *term* berdasarkan K dokumen disebut Kamus *term* dengan t*erm-term* dalam Kamus *term* diberi kode order *term* T_1, T_2, … T_s.

(e) Membuat Algoritma Tabel Munculnya *Term* dalam setiap kalimat (yang mengandung *term-term* hasil *parsing text*) dalam dokumen, dengan rincian tahapan sebagai berikut:

- *Term-term* hasil *parsing text* disusun di setiap kalimat yang sesuai dengan munculnya *term*.
- Mengambil *term* yang berada pada munculnya *term* pertama di setiap kalimat
- Membuat tabel munculnya *term* di setiap kalimat untuk masing-masing dokumen teks.

(f) Menghitung frekuensi munculnya *term* pertama dalam setiap kalimat untuk masing-masing dokumen teks.

(g) Membuat Algoritma Kumulatif Peluang Empiris Munculnya *Term* Pertama dalam koordinat Kartesius, dengan rincian tahapan sebagai berikut:

- Membuat koordinat Kartesius.
- Meletakkan frekuensi munculnya *term* pertama setiap kalimat dalam dokumen kesatu di koordinat Kartesius sebagai Kumulatif Peluang Empiris $\hat{F}(x)$,

- Meletakkan frekuensi munculnya *term* pertama setiap kalimat dalam dokumen kedua di koordinat Kartesius sebagai Kumulatif Peluang hipotesa $F(x)$.

(h) Mengimplementasikan uji K-S dalam munculnya *term* pertama antara dua dokumen untuk mengetahui apakah kedua dokumen teks tersebut mempunyai pola munculnya *term* pertama yang sama atau berbeda?

3.2 Tahapan Pola munculnya Pasangan *Term* Pertama

Tahapan pola munculnya pasangan *term* pertama adalah langkah-langkah untuk menjawab tujuan kedua yaitu membuat algoritma-algoritma untuk mengidentifikasi kesamaan pola dokumen melalui munculnya pasangan *term* pertama di setiap kalimat (yang mengandung *term-term* hasil *parsing text*) dalam dokumen dengan pendekatan jarak *Euclidean*, dengan rincian tahapan sebagai berikut:

(a) Membuat Algoritma Pasangan *Term*, dengan rincian tahapan sebagai berikut:
- Diasumsikan order munculnya *term-term* di masing-masing kalimat adalah berskala satu satuan.
- Munculnya *term* ke-1 dan *term* ke-2 di masing-masing kalimat di setiap dokumen akan berpasangan dan dinamakan sebagai pasangan *term*.
- Pasangan *term* tersebut disusun/ditabelkan sesuai dengan munculnya *term* disetiap kalimat di masing-masing dokumen sebagai tabel pasangan *term*.

(b) Membuat Algoritma Menghitung Jarak *Euclidean* munculnya pasangan *term* pertama dari kedua dokumen teks. Algoritma ini didasari oleh algoritma kamus *term* dan algoritma pasangan *term*.

(c) Membuat Algoritma Acuan Kesamaan Dokumen, dengan rincian tahapan sebagai berikut:
- Tabelkan semua jarak antara pasangan *term* dari dua dokumen teks.
- Mencari jarak maksimum (D = maks (d_i)). Jika jarak maksimum (D) tidak melebihi dari 3 maka kedua dokumen mempunyai pola munculnya pasangan *term* pertama sama, yang berarti kedua dokumen dikatakan sama.

Jarak maksimum tidak melebihi 3 satuan, dikarenakan munculnya pasangan *term* dan frekuensi munculnya pasangan *term* mempunyai skala satu satuan.

3.3 Tahapan Pola Munculnya Tiga *Term* Pertama

Tahapan pola munculnya tiga *term* pertama adalah langkah-langkah untuk menjawab tujuan ketiga yaitu membuat algoritma-algoritma dan teorema untuk mengidentifikasi kesamaan pola dokumen melalui munculnya tiga *term* pertama di setiap kalimat (yang mengandung *term-term* hasil *parsing text*) dalam dokumen dengan pendekatan *Bayesian Network* (BN), dengan rincian tahapan sebagai berikut:

(a) Membuat Algoritma Order Munculnya *Term*, dengan tahapan sebagai berikut:
- Mengambil *term* dari kamus *term* dan tempatkan *term* ke dalam order *term* setiap kalimat di masing-masing dokumen teks.
- Menandai order munculnya *term* setiap kalimat untuk masing-masing dokumen dengan nomor urut.

(b) Membuat Algoritma Pola Munculnya *Term* yang didasari oleh algoritma order munculnya *term* dengan tahapan sebagai berikut:
- Mengambil semua *term* yang tepat di setiap kalimat dari order munculnya *term* hasil algoritma order kemunculan *term*.
- Mengatur semua *term-term* mengikuti struktur setiap kalimat dalam setiap dokumen sebagai pola munculnya *term*.
- Mengambil tiga *term* pertama dari pola munculnya *term* di setiap kalimat dari setiap dokumen.
- Membangun distribusi gabungan dari pola munculnya tiga *term* pertama dalam setiap kalimat dari setiap dokumen.

(c) Membuat Algoritma *Likelihood* Munculnya Tiga *Term* Pertama, dengan tahapan sebagai berikut:
- Mengambil munculnya tiga *term* pertama dari setiap kalimat berdasarkan pola munculnya *term* hasil algoritma pola munculnya *term*.
- Mengumpulkan dan menghitung individu *term*

- Mengumpulkan dan menghitung frekuensi munculnya *term* pertama dari semua kalimat sebagai sebuah kelompok kemudian menghitung probabilitas setiap *term* antara munculnya *term* pertama dalam kelompok ini.
- Kumpulkan munculnya *term* kedua dan ketiga dari semua kalimat dan menghitung probabilitas dalam kelompok mereka seperti yang dilakukan untuk *term* pertama.

- Menghitung *likelihood* setiap kalimat dengan mengalikan semua probabilitas tiga munculnya *term* pertama dalam kalimat.

(d) Membuat Algoritma Rasio *Likelihood* Dokumen, dengan tahapan sebagai berikut:

- Menghitung *likelihood* setiap kalimat diwakili oleh munculnya tiga *term* yang pertama dalam setiap dokumen hasil algoritma pola kemunculan *term*.
- Menghitung *likelihood* setiap dokumen dengan mengalikan *likelihood* setiap kalimat dalam dokumen terkait.
- Menghitung rasio *likelihood* dua dokumen dengan membagi *likelihood* dokumen pilihan dengan dokumen yang lainnya.
- Menentukan perbandingan antara pasangan dokumen dengan menggunakan rasio *likelihood* masing-masing dokumen berdasarkan acuan standar nilai pembeda *Bayes Factor* (Kass dan Raftery, 1995).

BAB 4
POLA MUNCULNYA *TERM* PERTAMA

Tujuan utama bab ini adalah membahas pendeteksian kesamaan pola dokumen teks melalui munculnya *term* pertama di setiap kalimat dalam dokumen teks dengan uji *Kolmogorov-Smirnov* (uji K-S). Kalimat dalam dokumen teks adalah order *term* yang dihasilkan dari *parsing text* yang merupakan bagian dari proses *Latent Semantic Analysis* (LSA).

4.1 Implementasi *Parsing Text*

Implementasi *parsing text* menggunakan 6 dokumen uji yaitu Dok-1, Dok-2, Dok-3, Dok-4, Dok-5 dan Dok-6. Skenario untuk 6 dokumen sebagai berikut:

1. Dok-1 = Dok-2, hanya dirubah struktur kalimat dalam paragrafnya.
2. Dok-3 hampir sama dengan Dok-4, hanya ada perubahan struktur kata dalam kalimat
3. Dok-5 merupakan gabungan Dok-1 dan Dok-3.
4. Dok-6 merupakan dokumen yng dibuat berbeda dengan 5 dokumen lainnya.

Skenario dan karakteristik isi dari 6 dokumen untuk masing-masing dokumen terdapat dalam Tabel.4.1. Sedangkan isi masing-masing dokumen terdapat dalam Lampiran 1.

Tabel 4.1. Isi dokumen teks

Dokumen	Isi dokumen teks
Dok-1	berisi informasi atau kata-kata yang bermakna.
Dok-2	berisi sama dengan Dok-1, hanya dirubah struktur kalimat dalam paragrafnya.
Dok-3	memiliki tema yang sama dengan Dok-1, tetapi struktur kata dan konstruksi yang berbeda dengan Dok-1.
Dok-4	merupakan Dok-3 yang dilakukan perubahan struktur kata di dalam kalimat.
Dok-5	merupakan gabungan dari Dok-1 dan Dok-3.
Dok-6	dokumen teks yang berbeda dengan Dok-1, Dok-2, Dok-3, Dok-4 dan Dok-5.

4.2 Proses Pembentukan *Term* Dokumen

Pembentukan *term* diawali dengan pengambilan dokumen yang berisikan kalimat-kalimat yang terdiri dari beberapa kata untuk di *parsing text*. *Parsing text* mempunyai tiga langkah yaitu *Tokenizing*, *Filtering* dan *Stemming*. Sebagai ilustrasi diambil dokumen-1 (Dok-1) yang terdiri dari 91 kata (kata yang berulang dan tidak berulang) untuk di *parsing text* yang terdapat dalam Gambar 4.1, Gambar 4.2 dan Gambar 4.3. Untuk *parsing text* dokumen yang lainnya terdapat dalam lampiran 1.

> Akustik adalah ilmu yang mempelajari perilaku bunyi dan sangat penting pada ruangan. Dinding yang keras dan polos dari sebuah ruangan akan memantulkan bunyi dan membuat ruangan tersebut bergema. Ruangan yang kecil akan terbantu mencegah hal ini bila ada bahan pada dinding dan langit-langit yang menyerap bunyi. Tirai dan karpet yang tebal juga akan membantu. Pada ruangan yang besar seperti gedung konser, diperlukan permukaan yang halus dan keras di belakang para peminat atau penyanyi untuk membantu membawa bunyi ke arah penonton, dan bahan yang menyerap bunyi di belakang gedung untuk mencegah gema.

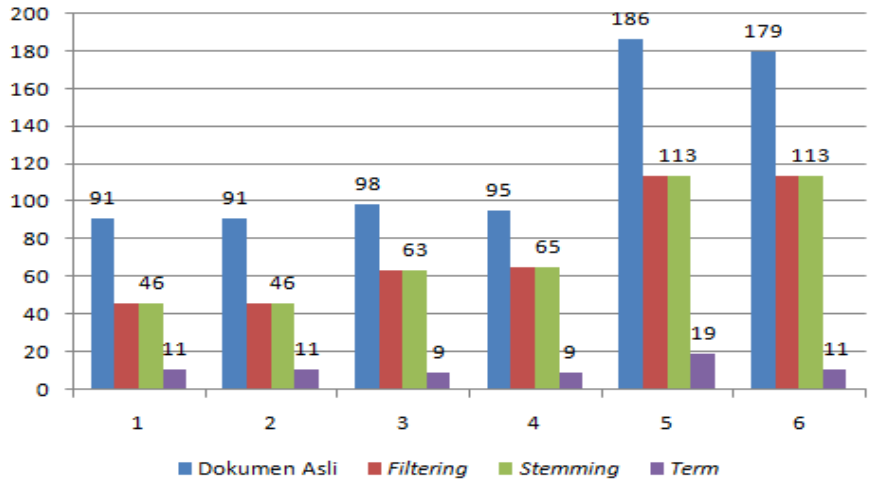
Gambar 4.1 Dokumen-1 Asli (total 91 kata)

> akustik ilmu mempelajari perilaku bunyi ruangan dinding keras polos ruangan memantulkan bunyi membuat ruangan bergema ruangan kecil mencegah bahan dinding langit-langit menyerap bunyi tirai karpet tebal ruangan besar gedung konser diperlukan permukaan halus keras peminat penyanyi membawa bunyi arah penonton bahan menyerap bunyi gedung mencegah gema

Gambar 4.2 Hasil *filtering* dokumen-1 (total 46 kata)

> akustik ilmu pelajari perilaku bunyi ruang dinding keras polos ruang pantul bunyi buat ruang gema ruang kecil cegah bahan dinding langit-langit serap bunyi tirai karpet tebal ruang besar gedung konser perlu muka halus keras minat penyanyi bawa bunyi arah penonton bahan serap bunyi gedung cegah gema

Gambar 4.3 Hasil *stemming* dokumen-1 (total 46 kata)



Gambar 4.4 Jumlah kata setelah proses *parsing text*

Proses *filtering* dan *stemming* dilakukan dengan cara yang sama untuk Dok-2, Dok-3, Dok-4, Dok-5 dan Dok-6 terdapat dalam lampiran. Jumlah kata Dok-1 setelah proses *filtering* dan *stemming* menjadi sebanyak 46 kata (kata yang berulang dan tidak berulang). *Term* diambil dari hasil proses *stemming* dengan banyak kata yang muncul minimal 2 kali, kecuali jika terdapat kata yang berada di masing-masing dokumen. *Term-term* yang terdapat dalam Dok-1 sebanyak 11 *term* yang terdiri dari ruang (5), bunyi (5), gema (2), cegah (2), gedung (2), serap (2), keras (2), langit (2), dinding (2), bahan (2), akustik (1). Dengan cara yang sama dilakukan untuk 6 dokumen menghasilkan *term-term* terdapat dalam Tabel 4.2 dan frekuensi munculnya *term* dalam Tabel 4.3.

Tabel 4.2. *Term-term* dalam setiap dokumen

Dokumen	**Jumlah Term**	***Term-term* dalam setiap dokumen**
Dok-1 & Dok-2	11	ruang, bunyi, gema, cegah, gedung, serap, keras, langit, dinding, bahan, akustik
Dok-3 & Dok-4	9	musik, budaya, lampung, festival, ada, hingga, tradisional, daerah, akustik
Dok-5	19	ruang, bunyi, gema, cegah, gedung, serap, keras, langit, dinding, musik, budaya, lampung, festival, ada, hingga, tradisional, daerah, bahan, akustik
Dok-6	9	negosiasi, dasar, ikut, kantor, rapat, orang, lebih, manusia, teknik, kalangan, bangun

Kode *term* didapatkan dari Tabel 4.3 yaitu T_1, T_2, …, T_{30} yang disusun dari jumlah frekuensi munculnya *term* terbesar hingga terkecil untuk masing-masing dokumen. Jika ada *term* yang sama diantara dokumen, maka urutan kode *term* disesuaikan dengan urutan dokumen yang mempunyai *term* terbanyak. Kode *term* disusun dalam Tabel 4.4.

Tabel 4.3. Frekuensi munculnya *term* semua dokumen teks

Term	**Frekuensi munculnya term di**					
	Dok-1	**Dok-2**	**Dok-3**	**Dok-4**	**Dok-5**	**Dok-6**
ruang	5	5	0	0	5	0
bunyi	5	5	0	0	5	0
gema	2	2	0	0	2	0
cegah	2	2	0	0	2	0
gedung	2	2	0	0	2	0
serap	2	2	0	0	2	0
keras	2	2	0	0	2	0
langit	2	2	0	0	2	0
dinding	2	2	0	0	2	0
musik	0	0	7	7	7	0
budaya	0	0	4	4	4	0
lampung	0	0	4	4	4	0
festival	0	0	3	3	3	0
ada	0	0	2	2	2	0
hingga	0	0	2	2	2	0
tradisional	0	0	2	2	2	0
daerah	0	0	1	2	2	0
bahan	2	2	0	0	2	0
akustik	1	1	1	1	2	0
negosiasi	0	0	0	0	0	8
lebih	0	0	0	0	0	4
dasar	0	0	0	0	0	3
ikut	0	0	0	0	0	2
kantor	0	0	0	0	0	2
rapat	0	0	0	0	0	2
orang	0	0	0	0	0	2
manusia	0	0	0	0	0	2
teknik	0	0	0	0	0	2
kalangan	0	0	0	0	0	2
bangun	0	0	0	0	0	2

Tabel 4.4. Kode *term* semua dokumen teks

Kode Term	Term	Frekuensi term					
		Dok-1	Dok-2	Dok-3	Dok-4	Dok-5	Dok-6
T_1	ruang	5	5	0	0	5	0
T_2	bunyi	5	5	0	0	5	0
T_3	gema	2	2	0	0	2	0
T_4	cegah	2	2	0	0	2	0
T_5	gedung	2	2	0	0	2	0
T_6	serap	2	2	0	0	2	0
T_7	keras	2	2	0	0	2	0
T_8	langit	2	2	0	0	2	0
T_9	dinding	2	2	0	0	2	0
T_{10}	musik	0	0	7	7	7	0
T_{11}	budaya	0	0	4	4	4	0
T_{12}	lampung	0	0	4	4	4	0
T_{13}	festival	0	0	3	3	3	0
T_{14}	ada	0	0	2	2	2	0
T_{15}	hingga	0	0	2	2	2	0
T_{16}	tradisional	0	0	2	2	2	0
T_{17}	daerah	0	0	1	2	2	0
T_{18}	bahan	2	2	0	0	2	0
T_{19}	akustik	1	1	1	1	2	0
T_{20}	negosiasi	0	0	0	0	0	8
T_{21}	lebih	0	0	0	0	0	4
T_{22}	dasar	0	0	0	0	0	3
T_{23}	ikut	0	0	0	0	0	2
T_{24}	kantor	0	0	0	0	0	2
T_{25}	rapat	0	0	0	0	0	2
T_{26}	orang	0	0	0	0	0	2
T_{27}	manusia	0	0	0	0	0	2
T_{28}	teknik	0	0	0	0	0	2
T_{29}	kalangan	0	0	0	0	0	2
T_{30}	bangun	0	0	0	0	0	2

4.3 Algoritma Kamus *Term*

Pembuatan kamus *term* ini sangatlah penting untuk algoritma berikutnya menuju pola dokumen teks berdasarkan munculnya *term*, diantaranya pola

munculnya *term* pertama sebuah dokumen. Tahapan pembuatan algoritma tentang penyusunan Kamus *Term* dirinci dalam Algoritma 4.1.

Algoritma 4.1: Pembuatan Kamus *Term*

Langkah 1. Siapkan sebanyak K dokumen akan diuji kesamaannya.

Langkah 2. Namakan masing-masing dokumen terurut dari dokumen 1 (Dok-1) sampai dengan dokumen K (Dok-K).

Langkah 3. Proseslah semua K dokumen dengan proses *parsing text* untuk menghasilkan *term-term* $\{T_1, T_2, \ldots, T_s\}$ dari setiap kalimat yang berada di masing-masing dokumen.

Langkah 4. Membuat daftar order *term* berdasarkan K dokumen disebut Kamus *term*, dengan *term-term* dalam Kamus *term* diberi kode order *term* $T_1, T_2, \ldots T_s$.

Tabel 4.5. Kamus *Term*

Kode *term*	***Term***	**Kode *term***	***Term***	**Kode *term***	***Term***
T_1	ruang	T_{11}	budaya	T_{21}	lebih
T_2	bunyi	T_{12}	lampung	T_{22}	dasar
T_3	gema	T_{13}	festival	T_{23}	ikut
T_4	cegah	T_{14}	ada	T_{24}	kantor
T_5	gedung	T_{15}	hingga	T_{25}	rapat
T_6	serap	T_{16}	tradisional	T_{26}	orang
T_7	keras	T_{17}	daerah	T_{27}	manusia
T_8	langit	T_{18}	bahan	T_{28}	teknik
T_9	dinding	T_{19}	akustik	T_{29}	kalangan
T_{10}	musik	T_{20}	negosiasi	T_{30}	bangun

4.4 Pola Munculnya *Term* Pertama

Penulis suatu dokumen untuk merepresentasikan idenya ke dalam kalimat dapat dibedakan pada pemilihan kata-katanya. Rentetan kata yang terpilih tersebut akan membangun kesatuan ide penulis yang konvergen. Munculnya *term* pertama di suatu kalimat cenderung bisa membedakan pola penyampaian ide antar kalimat. Pembuat algoritma pembuatan tabel munculnya *term* pertama setiap kalimat dalam dokumen teks, dirinci dalam Algoritma 4.2.

Algoritma 4.2: Pembuatan Tabel Munculnya *Term* Pertama

Langkah 1. *Term-term* hasil *parsing text* $\{T_1, T_2, \ldots, T_s\}$ disusun di setiap kalimat yang sesuai dengan munculnya *term*,

Langkah 2. Ambil *term* yang berada pada munculnya *term* pertama $\forall T_1 \in K$, di setiap kalimat dalam K dokumen,

Langkah 3. Membuat tabel munculnya *term* pertama di setiap kalimat untuk masing-masing dokumen teks.

Pola dokumen teks yang dimaksud merupakan bentuk atau model yang bisa dipakai untuk mengidentifikasi kesamaan suatu dokumen. Pola dokumen teks yang diidentifikasi adalah pola munculnya *term* pertama dalam setiap dokumen teks. Untuk mengidentifikasi pola munculnya *term* pertama, terlebih dahulu membuat tabel *term* setiap kalimat semua dokumen teks yang terdapat dalam Tabel 4.6. Frekuensi munculnya *term* pertama setiap dokumen teks terdapat dalam Tabel 4.7.

Tabel 4.6. *Term* setiap kalimat semua dokumen

	Kalimat	**Munculnya *term* ke-**									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Dok-1	1	T_{19}	T_2	T_1							
	2	T_9	T_7	T_1	T_2	T_1	T_3				
	3	T_1	T_4	T_{18}	T_9	T_8	T_8	T_6	T_2		
	4	T_1	T_5	T_7	T_2	T_{18}	T_6	T_2	T_5	T_4	T_3
Dok-2	**Kalimat**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	1	T_1	T_4	T_{18}	T_9	T_8	T_8	T_6	T_2		
	2	T_1	T_5	T_7	T_2	T_{18}	T_6	T_2	T_5	T_4	T_3
	3	T_{19}	T_2	T_1							
	4	T_9	T_7	T_1	T_2	T_1	T_3				
Dok-3	**Kalimat**	1	2	3	4	5	6	7	8	9	10
	1	T_{17}	T_{12}	T_{10}	T_{16}	T_{15}	T_{11}	T_{10}			
	2	T_{10}	T_{15}	T_{12}							
	3	T_{10}	T_{10}	T_{19}							
	4	T_{10}	T_{11}	T_{11}							
	5	T_{13}	T_{14}	T_{11}	T_{10}	T_{16}					
	6	T_{13}	T_{13}	T_{14}	T_{12}	T_{12}					

	Kalimat	Munculnya *term* ke-									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Dok-4	1	T_{10}	T_{17}	T_{12}	T_{17}	T_{16}	T_{15}	T_{11}	T_{10}		
	2	T_{12}	T_{10}	T_{15}							
	3	T_{10}	T_{19}	T_{10}							
	4	T_{10}	T_{11}	T_{11}							
	5	T_{11}	T_{10}	T_{16}	T_{14}	T_{13}					
	6	T_{12}	T_{14}	T_{13}	T_{12}	T_{13}					
	Kalimat	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Dok-5	1	T_{10}	T_{17}	T_{12}	T_{17}	T_{16}	T_{15}	T_{11}	T_{10}		
	2	T_{12}	T_{10}	T_{15}							
	3	T_{10}	T_{19}	T_{10}							
	4	T_{10}	T_{11}	T_{11}							
	5	T_{11}	T_{10}	T_{16}	T_{14}	T_{13}					
	6	T_{12}	T_{14}	T_{13}	T_{12}	T_{13}					
	7	T_{19}	T_{2}	T_{1}							
	8	T_{9}	T_{7}	T_{1}	T_{2}	T_{1}	T_{3}				
	9	T_{1}	T_{4}	T_{18}	T_{9}	T_{8}	T_{8}	T_{6}	T_{2}		
	10	T_{1}	T_{5}	T_{7}	T_{2}	T_{18}	T_{6}	T_{2}	T_{5}	T_{4}	T_{3}
	Kalimat	1	2	3	4	5	6	7	8	9	10
Dok-6	1	T_{30}	T_{21}	T_{21}							
	2	T_{24}	T_{20}	T_{25}	T_{25}	T_{20}					
	3	T_{23}	T_{23}	T_{26}							
	4	T_{20}	T_{27}	T_{27}							
	5	T_{26}	T_{29}	T_{29}							
	6	T_{22}	T_{22}	T_{28}	T_{28}	T_{20}	T_{20}	T_{30}	T_{21}	T_{21}	

4.5 Pembuatan Kumulatif Peluang Empiris Munculnya *Term* Pertama

Algoritma pembuatan kumulatif peluang empiris munculnya *term* pertama setiap kalimat dalam dokumen teks, dirinci dalam Algoritma 4.3.

Algoritma 4.3: Kumulatif peluang empiris munculnya *term* pertama

Langkah 1. Buat koordinat Kartesius (*X*, *Y*, *Z*)

Langkah 2. Letakkan frekuensi munculnya *term* pertama setiap kalimat dalam dokumen kesatu di koordinat Kartesius sebagai kumulatif peluang empiris $\hat{F}(x)$,

Langkah 3. Letakkan frekuensi munculnya *term* pertama setiap kalimat dalam dokumen kedua di koordinat Kartesius sebagai kumulatif peluang hipotesa $F(x)$.

Tabel 4.7. Frekuensi munculnya *term* pertama setiap dokumen teks

Term	**Kode Term**	**Dok-1**	**Dok-2**	**Dok-3**	**Dok-4**	**Dok-5**	**Dok-6**
ruang	T_1	2	2	0	0	2	0
dinding	T_9	1	1	0	0	1	0
musik	T_{10}	0	0	3	3	3	0
budaya	T_{11}	0	0	0	1	1	0
lampung	T_{12}	0	0	0	2	2	0
festival	T_{13}	0	0	2	0	0	0
daerah	T_{17}	0	0	1	0	0	0
akustik	T_{19}	1	1	0	0	1	0
negosiasi	T_{20}	0	0	0	0	0	1
dasar	T_{22}	0	0	0	0	0	1
ikut	T_{23}	0	0	0	0	0	1
kantor	T_{24}	0	0	0	0	0	1
orang	T_{26}	0	0	0	0	0	1
bangun	T_{30}	0	0	0	0	0	1

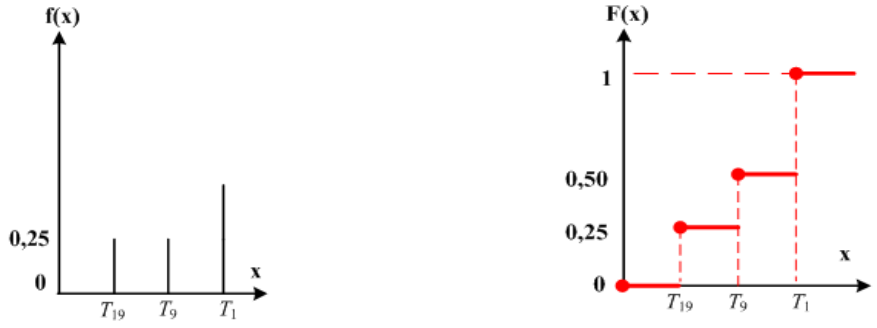
Tabel 4.8. Munculnya *term* pertama untuk Dok-1

Kalimat	**Dok-1**	**Frekuensi**
1	T_{19}	1
2	T_9	1
3	T_1	2

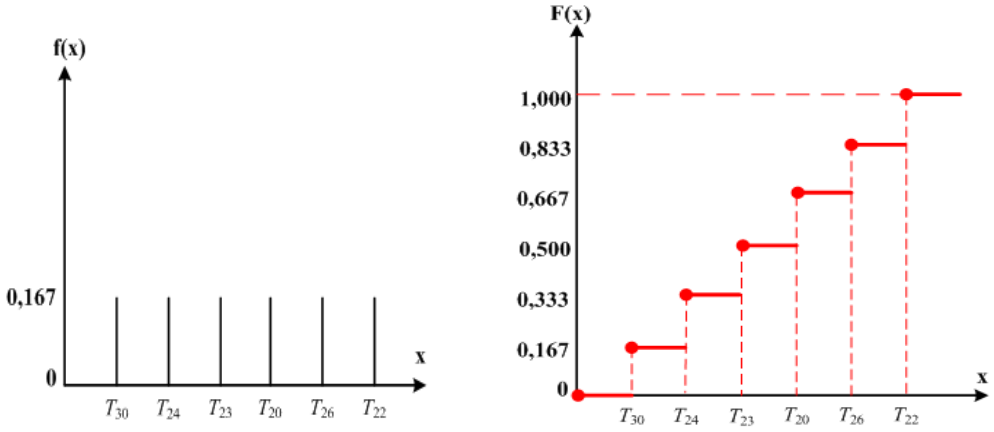
Tabel 4.9. Munculnya *term* pertama untuk Dok-1 dan Dok-6

Kalimat	**Dok-1**	**Frekuensi**	**Dok-6**	**Frekuensi**
1	T_{19}	1	T_{30}	1
2	T_9	1	T_{24}	1
3	T_1	2	T_{23}	1
4			T_{20}	1
5			T_{26}	1
6			T_{22}	1

Berdasarkan Algoritma 4.3 tentang kumulatif peluang empiris munculnya *term* pertama dalam koordinat Kartesius, maka dapat digambarkan pola peluang dan kumulatif peluang munculnya *term* pertama untuk Dok-1 (Tabel 4.8) dalam Gambar 4.5, sedangkan untuk pola peluang dan kumulatif peluang munculnya *term* pertama Dok-6 terdapat dalam Gambar 4.6.



Gambar 4.5 Pola peluang dan kumulatif peluang munculnya *term* pertama Dok-1



Gambar 4.6 Pola peluang dan kumulatif peluang munculnya *term* pertama Dok-6

4.6 Perhitungan Uji K-S

Pembedaan dokumen dengan menggunakan informasi munculnya *term* pertama dalam suatu kalimat akan digunakan perhitungan jarak D_q antara munculnya *term* pertama antar dua dokumen. Penghitungan jarak D_q menggunakan uji K-S seperti yang telah dibahas dalam sub bab 2.3 dalam persamaan (2.4) dengan $D_q = \max\left|\hat{F}_A(x) - F_B(x)\right|$.

Perhitungan kumulatif peluang (KP) diskrit untuk nilai x_1, x_2, …, x_i dengan probabilitas $p_i = P(x_i)$ sebagai berikut: (Sahoo, 2013)

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} p(x_i). \qquad (4.1)$$

Perhitungan Peluang dan kumulatif peluang dengan persamaan (4.1) untuk Dok-1 dan Dok-2 terdapat dalam Tabel 4.10, sedangkan untuk Dok-1 dan Dok-6 terdapat dalam Tabel 4.11.

Tabel 4.10. Peluang Munculnya *Term* Pertama, Kumulatif Peluang (KP) Munculnya *Term* Pertama dan D_q untuk Dok-1 dan Dok-2

Kalimat	Dok-1	Peluang	KP D-1	Dok-2	Peluang	KP D-2	D_q
1	T_{19}	0,250	0,250	T_1	0,500	0,500	0,250
2	T_9	0,250	0,500	T_{19}	0,250	0,750	0,250
3	T_1	0,500	1,000	T_9	0,250	1,000	0,000

Tabel 4.11. Peluang Munculnya *Term* Pertama, Kumulatif Peluang (KP) Munculnya *Term* Pertama dan D_q untuk Dok-1 dan Dok-6

Kalimat	Dok-1	Peluang	KP D-1	Dok-6	Peluang	KP D-6	D_q
1	T_{19}	0,250	0,250	T_{30}	0,167	0,167	0,083
2	T_9	0,250	0,500	T_{24}	0,167	0,333	0,167
3	T_1	0,500	1,000	T_{23}	0,167	0,500	0,500
4			1,000	T_{20}	0,167	0,667	0,333
5			1,000	T_{26}	0,167	0,833	0,167
6			1,000	T_{22}	0,167	1,000	0,000

Dengan perhitungan yang sama seperti Tabel 4.10 dan Tabel 4.11, perhitungan D_q untuk dokumen lainnya terdapat dalam Tabel 4.12.

Tabel 4.12. Hasil perhitungan D_q

	Dok-1	Dok-2	Dok-3	Dok-4	Dok-5	Dok-6
Dok-1	0,000	0,250	0,167	0,333	0,400	0,500
Dok-2	0,250	0.000	0,167	0,333	0,400	0,500
Dok-3	0,167	0,167	0.000	0,333	0,400	0,500
Dok-4	0,333	0,333	0,333	0.000	0,400	0,500
Dok-5	0,400	0,400	0,400	0,400	0.000	0,167
Dok-6	0,500	0,500	0,500	0,500	0,167	0.000

Hipotesis yang digunakan dalam perhitungan uji K-S sebagai berikut:

H_0: dua dokumen teks memiliki pola munculnya *term* pertama yang sama.

H_1: dua dokumen teks tidak memiliki pola munculnya *term* pertama yang sama.

Perhitungan uji K-S dengan merujuk Tabel 2.1 dan Tabel 4.10 untuk Dok-1 dan Dok-2 diperoleh D_q = 0,250, q = 8 (banyaknya *term* yang muncul pertama dalam Dok-1 dan Dok-2) sebagai berikut:

$\left(\sqrt{q} + 0{,}12 + \frac{0{,}11}{\sqrt{q}}\right) D_q = \left(\sqrt{8} + 0{,}12 + \frac{0{,}11}{\sqrt{8}}\right) 0{,}250 = 0{,}747$, untuk α = 10 % didapatkan $c_{1-\alpha}$ = 1,2240. Sehingga $\left(\sqrt{q} + 0{,}12 + \frac{0{,}11}{\sqrt{q}}\right) D_q < c_{1-\alpha}$, yang dapat disimpulkan gagal menolak H_0 (Dok-1 dan Dok-2 memiliki pola munculnya *term* pertama yang sama).

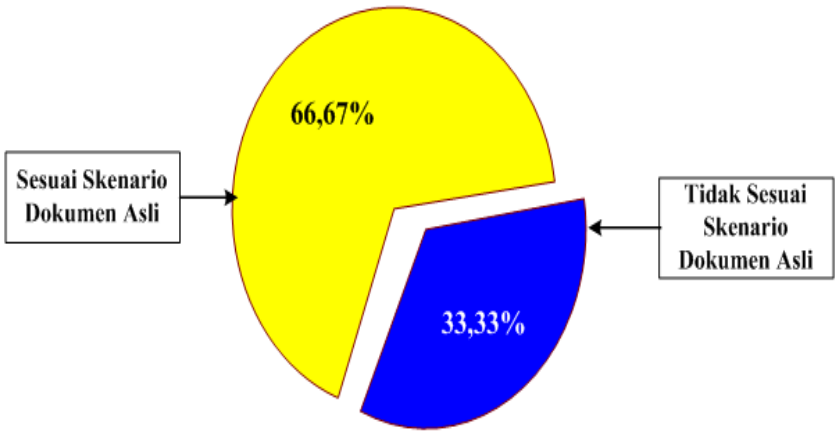
Dengan perhitungan yang sama, maka didapatkan hasil perhitungan uji K-S untuk 6 dokumen tertuang dalam Tabel 4.13.

Tabel 4.13. Hasil perhitungan uji K-S untuk 6 dokumen

No	Dokumen		Keputusan Uji K-S		Skenario Dokumen	
			Gagal Menolak H_0	Menolak H_0	Sesuai	Tidak Sesuai
1	Dok-1	Dok-2	√		√	
2		Dok-3	√			√
3		Dok-4	√			√
4		Dok-5		√	√	
5		Dok-6		√	√	
6	Dok-2	Dok-3	√			√
7		Dok-4	√			√
8		Dok-5		√	√	
9		Dok-6		√	√	
10	Dok-3	Dok-4	√		√	
11		Dok-5		√	√	
12		Dok-6		√	√	
13	Dok-4	Dok-5		√	√	
14		Dok-6		√	√	
15	Dok-5	Dok-6	√			√

Identifikasi dan kesamaan dokumen teks dengan memperhatikan pola munculnya *term* pertama setiap dua dokumen dari 6 dokumen uji, terdapat 15 pasangan dokumen. Hasil perhitungan uji *Kolmogorov-Smirnov* (uji K-S) untuk 15 pasangan dokumen yang diuji terdapat 10 pasangan dokumen (66,67 %) yang sesuai dengan skenario enam dokumen asli, terdiri dari:

a. Gagal menolak H_0 untuk pasangan:
   (Dok-1,Dok-2) dan (Dok-3,Dok-4).
b. Menolak H_0 untuk pasangan:
   (Dok-1,Dok-5), (Dok-1,Dok-6), (Dok-2, Dok-5), (Dok-2,Dok-6),
   (Dok-3,Dok-5), (Dok-3,Dok-6), (Dok-4,Dok-5), dan (Dok-4,Dok-6)



Gambar 4.7 Perbandingan hasil uji K-S dengan Skenario dokumen asli

BAB 5
POLA MUNCULNYA PASANGAN *TERM* PERTAMA

Tujuan utama bab ini adalah membahas identifikasi pola dokumen teks berdasarkan munculnya pasangan *term* pertama sebagai tindak lanjut penelitian Bab 4 dengan menggunakan hasil *parsing text* bagian dari *Latent Semantic Analysis* (LSA) terhadap dokumen dan jarak pasangan *term* antara dua dokumen. Masing-masing dokumen mempunyai *term* yang kemunculannya diurutkan dari munculnya *term* ke-1, ke-2 dan seterusnya. Munculnya *term* ke-1 dan ke-2 akan menjadi pasangan *term* di masing-masing dokumen teks. Pasangan *term* dan frekuensi kemunculan digambarkan dalam koordinat kartesian 3 dimensi (sumbu *X*, *Y* dan *Z*) yang membentuk titik. Titik-titik pasangan *term* di masing-masing dokumen dapat dihitung jaraknya antara dua dokumen teks.

Dokumen yang diuji dalam penelitian ini sebanyak 6 dokumen teks yang terdapat dalam Tabel 4.1. Dokumen yang telah diproses *parsing text* didapatkan *term-term* untuk masing-masing dokumen teks terdapat dalam Tabel 4.2. Pola dokumen yang diidentifikasi adalah pola munculnya pasangan *term* pertama dalam setiap dokumen. Untuk mendapatkan pola dokumen teks dengan menggunakan hasil *parsing text* dan jarak *term* antara dua dokumen, perlu dibuat algoritma dan teorema yang terdiri dari Algoritma 5.1 tentang Pasangan *Term*, Teorema 5.1 tentang Penghitungan jarak 2 titik-titik pasangan *term* dokumen teks dan Algoritma 5.2 tentang Acuan kesamaan dokumen teks. Adapun tahapan uraian algoritma dan teorema ini sebagai berikut:

5.1 Algoritma Pasangan *Term*

Dalam sebuah kalimat munculnya *term-term* akan berurutan di masing-masing kalimat dalam sebuah dokumen. Algoritma tentang pasangan *term* diperlukan untuk pembentukan algoritma-algoritma. Pembuatan algoritma tentang pasangan *term* dirinci dalam Algoritma 5.1.

Algoritma 5.1: Pembuatan Pasangan *Term*

Langkah 1. Asumsikan order munculnya *term-term* {T_1, T_2, …, T_n} di masing-masing kalimat adalah berjarak satu satuan $d_{|T_1T_2|} = 1$,

Langkah 2. Munculnya *term* ke-1 dan *term* ke-2 di masing-masing kalimat di setiap dokumen berpasangan dan dinamakan sebagai pasangan *term.*

Langkah 3. Menyusun pasangan *term* dalam tabel sesuai munculnya *term* disetiap kalimat di masing-masing dokumen sebagai tabel pasangan *term*.

Implementasi Algoritma 5.1, kedalam enam dokumen uji diperoleh tabel pasangan *term* yang ditunjukkan dalam Tabel 5.1A sampai dengan Tabel 5.1F di bawah ini.

Tabel 5.1A. Pasangan *Term* Dok-1

Dok-1	Munculnya *term* ke-	
Kalimat	1	2
1	T_{19}	T_2
2	T_9	T_7
3	T_1	T_4
4	T_1	T_5

Tabel 5.1B. Pasangan *Term* Dok-2

Dok-2	Munculnya *term* ke-	
Kalimat	1	2
1	T_1	T_4
2	T_1	T_5
3	T_{19}	T_2
4	T_9	T_7

Tabel 5.1C. Pasangan *Term* Dok-3

Dok-3	Munculnya *term* ke-	
Kalimat	1	2
1	T_{17}	T_{12}
2	T_{10}	T_{15}
3	T_{10}	T_{10}
4	T_{10}	T_{11}
5	T_{13}	T_{14}
6	T_{13}	T_{13}

Tabel 5.1D. Pasangan *Term* Dok-4

Dok-4	Munculnya *term* ke-	
Kalimat	1	2
1	T_{10}	T_{17}
2	T_{12}	T_{10}
3	T_{10}	T_{19}
4	T_{10}	T_{11}
5	T_{11}	T_{10}
6	T_{12}	T_{14}

Tabel 5.1E. Pasangan *Term* Dok-5

Dok-5	Munculnya *term* ke-	
Kalimat	1	2
1	T_{10}	T_{17}
2	T_{12}	T_{10}
3	T_{10}	T_{19}
4	T_{10}	T_{11}
5	T_{11}	T_{10}
6	T_{12}	T_{14}
7	T_{19}	T_2
8	T_9	T_7
9	T_1	T_4
10	T_1	T_5

Tabel 5.1F. Pasangan *Term* Dok-6

Dok-6	Munculnya *term* ke-	
Kalimat	1	2
1	T_{30}	T_{21}
2	T_{24}	T_{20}
3	T_{20}	T_{23}
4	T_{20}	T_{27}
5	T_{26}	T_{29}
6	T_{22}	T_{22}

Pola munculnya pasangan *term* dari dua dokumen teks diperoleh dengan menggunakan rumus jarak (*d*) yang terdapat persamaan (2.5) antara titik-titik pasangan *term* (*term* ke-1 dan *term* ke-2).

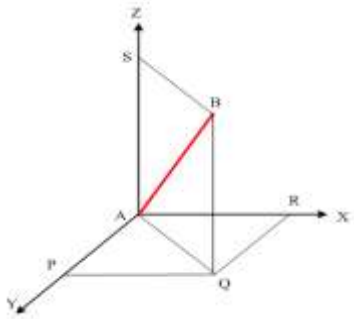
5.2 Algoritma Menghitung Jarak *Term*

Setelah pembuatan Algoritma 5.1 tentang Pasangan *Term*, dilanjutkan dengan pembuatan algoritma menghitung jarak *term* kedua dokumen yang dituangkan dalam Algoritma 5.2. Sebagai ilustrasi menuju Algoritma 5.2, perhatikan Gambar 5.1 yang merupakan perhitungan $d_{|AB|}$ (jarak titik A dan titik B). Misalkan titik $A(X_A, Y_A, Z_A)$ dan titik $B(X_B, Y_B, Z_B)$, dalam segitiga APQ berlaku rumus *Phytagoras* $AQ^2 = AR^2 + AP^2$,

$$AQ^2 = (X_R - X_A)^2 + (Y_P - Y_A)^2.$$

Untuk segitiga ABQ berlaku $AB^2 = AQ^2 + AS^2$,

$$AB^2 = (X_R - X_A)^2 + (Y_P - Y_A)^2 + (Z_S - Z_A)^2$$



Gambar 5.1 Ilustrasi jarak $d_{|AB|}$

Algoritma 5.2: Menghitung Jarak Term

Langkah 1. Mengambil pasangan *term* munculnya *term* ke-1 dan *term* ke-2 (*term* ke-1, *term* ke-2) disetiap kalimat di masing-masing dokumen.

Langkah 2. Meletakkan munculnya *term* ke-1 disetiap kalimat untuk dokumen ke-*i* pada sumbu *X* dengan langkah satu satuan.

Langkah 3. Meletakkan munculnya *term* ke-2 disetiap kalimat untuk dokumen ke-*j* pada sumbu *Y* dengan langkah satu satuan.

Langkah 4. Meletakkan banyaknya frekuensi munculnya pasangan *term* ke-1 dan *term* ke-2 pada sumbu *Z*. Setiap pasangan *term* disetiap kalimat dari kalimat ke-1, ke-2, …, ke-i digambarkan dengan penambahan satu langkah ke atas setiap pasangan *term* pada sumbu *Z*. Koordinat dokumen ke-2 mengikuti koordinat dokumen ke-1.

Langkah 5. Menghitung jarak dari pasangan *term* untuk kedua dokumen dengan persamaan (2.5):

$$d_{|AB|} = \sqrt{(X_B - X_A)^2 + (Y_B - Y_A)^2 + (Z_B - Z_A)^2}$$

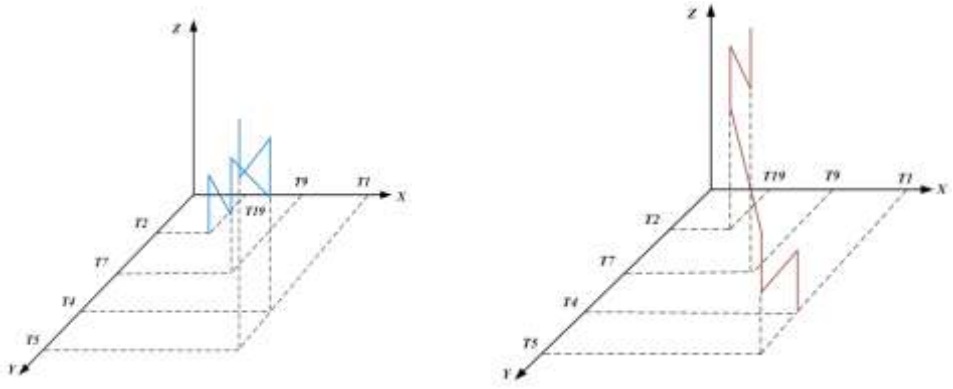
Ilustrasi dari Algoritma 5.2 untuk Dok-1 sebagai berikut:

- Terdapat pasangan *term* (T_{19},T_2), (T_9,T_7), (T_1,T_4) dan (T_1,T_5).
- Pada sumbu *X* terdapat T_{19}, T_9 dan T_1.
- Pada sumbu *Y* terdapat T_2, T_7, T_4 dan T_5.
- Pada sumbu *Z* terdapat (T_{19},T_2,1), (T_9,T_7,2), (T_1,T_4,3) dan (T_1,T_5,4), ada kenaikan satu langkah keatas setiap pasangan.

Ilustrasi dari Algoritma 5.2 untuk Dok-2 sebagai berikut:

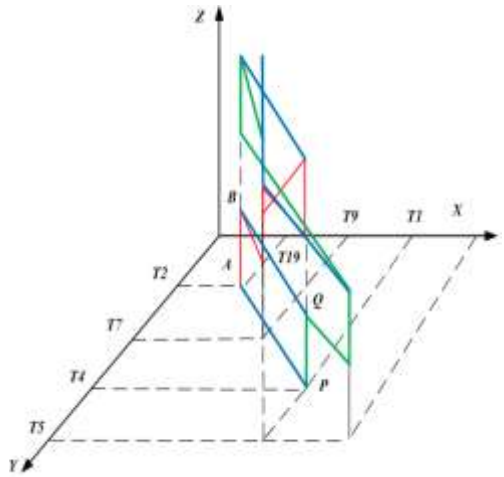
- Terdapat pasangan *term* (T_1,T_4), (T_1,T_5), (T_{19},T_2) dan (T_9,T_7).
- Pada sumbu *X* terdapat T_1, T_{19} dan T_9.
- Pada sumbu *Y* terdapat T_4, T_5, T_2 dan T_7.
- Pada sumbu *Z* terdapat (T_1,T_4,1), (T_1,T_5,2), (T_{19},T_2,3) dan (T_9,T_7,4), ada kenaikan satu langkah keatas setiap pasangan. Koordinat Dok-2 mengikuti koordinat Dok-1.

Aplikasi dari Algoritma 5.2 dapat dilihat dalam Gambar 5.3. Pola munculnya pasangan *term* pertama untuk dokumen kesatu (Dok-1) dan dokumen kedua (Dok-2) digambarkan dalam Gambar 5.2. Untuk jarak (*d*) pasangan *term* dari kedua dokumen dengan merujuk sub bab 2.4 tentang jarak *Euclidean*, dapat dilihat dalam Gambar 5.3. Gambar 5.3 menggambarkan jarak *d* pasangan *term* untuk Dok-1 dan Dok-2 yang perhitungan jarak (*d*) pasangan *term* kedua dokumen terdapat dalam Tabel 5.1A. Perhitungan jarak (*d*) untuk dokumen-dokumen yang lainnya terdapat dalam Tabel 5.1B, Tabel 5.1C, Tabel 5.1D, Tabel 5.1E dan Tabel 5.1F.



(a) (b)

Gambar 5.2 Pola munculnya pasangan *term* pertama (a). Dok-1 dan (b). Dok-2



Gambar 5.3 Jarak (*d*) pasangan *term* Dok-1 dan Dok-2

Perhitungan jarak untuk setiap munculnya pasangan *term* masing-masing dokumen dituangkan dalam Tabel 5.2A sampai dengan Tabel 5.2D yang diperoleh dari persamaan (2.5):

$$d_{|AB|} = \sqrt{(x_A - x_P)^2 + (y_A - y_P)^2 + (z_A - z_P)^2}$$

Misal untuk Dok-1 titik A(1,1,1) dengan dan pada Dok-2 titik B(3,3,1) maka

$$d_{|AB|} = \sqrt{(1-3)^2 + (1-3)^2 + (1-1)^2} = 2{,}83$$

Tabel 5.2A. Jarak (*d*) pasangan *term* munculnya pertama dari Dok-1 dan Dok-2

Dok-1					Dok-2					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{19}	1	T_2	1	1	T_1	3	T_4	3	1	2.83
T_9	2	T_7	2	2	T_1	3	T_5	4	2	2.24
T_1	3	T_4	3	3	T_{19}	1	T_2	1	3	2.83
T_1	3	T_5	4	4	T_9	2	T_7	2	4	2.24

Keterangan: Jarak (*d*) pasangan *term* munculnya pertama dari Dok-1 dan Dok-2 semuanya $d \leq 3$ (100%).

Tabel 5.2B. Jarak (*d*) pasangan *term* munculnya pertama dari Dok-3 dan Dok-4

Dok-3					Dok-4					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{17}	1	T_{12}	1	1	T_{10}	2	T_{17}	7	1	6.08
T_{10}	2	T_{15}	2	2	T_{12}	4	T_{10}	3	2	2.24
T_{10}	2	T_{10}	3	3	T_{10}	2	T_{19}	8	3	5.00
T_{10}	2	T_{11}	4	4	T_{10}	2	T_{11}	4	4	0.00
T_{13}	3	T_{14}	5	5	T_{11}	5	T_{10}	3	5	2.83
T_{13}	3	T_{13}	6	6	T_{12}	4	T_{14}	5	6	1.41

Keterangan: Jarak (*d*) pasangan *term* munculnya pertama dari Dok-3 dan Dok-4 yang berjarak $d \leq 3$ sebanyak 4 pasangan *term* (66,67%) dan jarak $d > 3$ sebanyak 2 pasangan *term* (33,33%).

Tabel 5.2C. Jarak (*d*) pasangan *term* munculnya pertama dari Dok-1 dan Dok-6.

Dok-1					Dok-6					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{19}	1	T_2	1	1	T_{30}	4	T_{21}	5	1	5.00
T_9	2	T_7	2	2	T_{24}	5	T_{20}	6	2	5.00
T_1	3	T_4	3	3	T_{20}	6	T_{23}	7	3	5.00
T_1	3	T_5	4	4	T_{20}	6	T_{27}	8	4	5.00
T_1	3	T_5	4	5	T_{26}	7	T_{29}	9	5	6.40
T_1	3	T_5	4	6	T_{22}	8	T_{22}	10	6	7.81

Keterangan: Jarak (*d*) pasangan *term* munculnya pertama dari Dok-1 dan Dok-6 semua berjarak $d > 3$ (100%).

Dengan cara yang sama dapat ditentukan jarak (*d*) untuk setiap pasangan *term* dari setiap dua dokumen sebanyak 15 pasangan dokumen. Seluruh hasil perhitungan jarak (*d*) terdapat dalam lampiran 3.

5.3 Algoritma Acuan Kesamaan Dokumen

Pembentukan algoritma tentang acuan kesamaan dua dokumen diperlukan sebagai hasil akhir untuk mengambil keputusan tentang dua dokumen mempunyai pola yang sama atau berbeda. Pembuatan algoritma tentang acuan kesamaan dua dokumen teks dirinci dalam Algoritma 5.3.

Algoritma 5.3: Acuan Kesamaan Dokumen

Langkah 1. Mengambil nilai jarak (*d*) dari perhitungan jarak setiap pasangan *term* dari setiap pasangan dokumen.

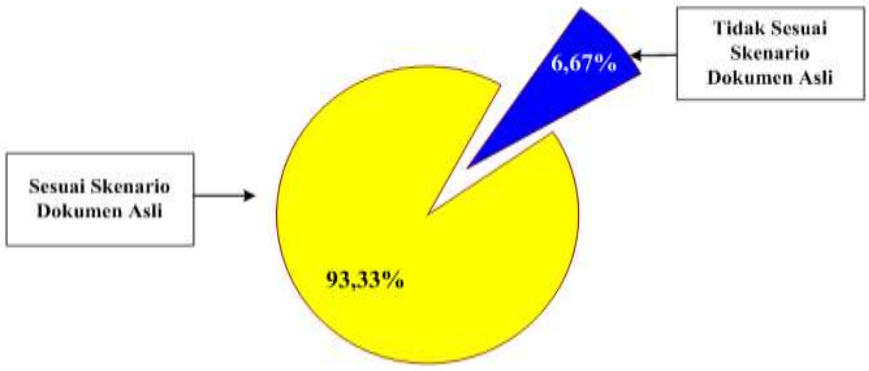
Langkah 2. Mengelompokkan $d \leq 3$ dan $d > 3$. Angka 3 merupakan perhitungan langkah satu satuan di koordinat (*X*,*Y*,*Z*) dari munculnya pasangan *term* dan frekuensi munculnya pasangan *term* di koordinat (*X*,*Y*,*Z*).

Langkah 3. Mengambil kesimpulan, jika semua jarak pasangan *term* bernilai $d \leq 3$, maka kedua dokumen mempunyai pola munculnya pasangan *term* pertama kedua dokumen sama. Jika semua jarak pasangan *term* bernilai $d > 3$, maka kedua dokumen mempunyai pola munculnya pasangan *term* pertama kedua dokumen tidak sama.

Tabel 5.3. Prosentase jarak (*d*) kedua dokumen

No	Dokumen		Prosentase (%)	
			$d \leq 3$	$d > 3$
1	Dok-1	Dok-2	100.00	
2	Dok-1	Dok-3		100.00
3	Dok-1	Dok-4		100.00
4	Dok-1	Dok-5	30.00	70.00
5	Dok-1	Dok-6		100.00
6	Dok-2	Dok-3		100.00
7	Dok-2	Dok-4		100.00
8	Dok-2	Dok-5	30.00	70.00
9	Dok-2	Dok-6		100.00
10	Dok-3	Dok-4	66.67	33.33
11	Dok-3	Dok-5	20.00	80.00
12	Dok-3	Dok-6		100.00
13	Dok-4	Dok-5	60.00	40.00
14	Dok-4	Dok-6		100.00
15	Dok-5	Dok-6		100.00

Identifikasi dan kesamaan dokumen teks dengan memperhatikan pola munculnya pasangan *term* pertama dari kedua dokumen yang diskenario (ada enam dokumen asli) dengan kombinasi pasangan 2 dokumen, hasilnya hanya satu pasangan dokumen (Dok-3 dan Dok-4) yang tidak sesuai dengan skenario dokumen asli (6,67 %).



Gambar 5.4 Perbandingan hasil jarak pasangan *term* antara dua dokumen dengan skenario dokumen asli

BAB 6
IDENTIFIKASI POLA STRUKTUR *TERM* DALAM DOKUMEN TEKS MENGGUNAKAN *BAYESIAN NETWORK*

Tujuan utama bab ini melanjutkan penelitian Bab 5 menyangkut pola dokumen teks berdasarkan identifikasi pola struktur *term* dalam dokumen teks menggunakan *Bayesian Network* (BN), untuk munculnya tiga *term* pertama dalam kalimat. Komunikasi seseorang sangat dipengaruhi oleh bahasa ibu, yang dapat mempengaruhi pola dokumen teks. Bagaimana seseorang bisa mengenali pola dokumen teks? Soehardjoepri dkk. (2013) telah berhasil mengembangkan identifikasi pola dokumen teks dari kemunculan *term* pertama dalam dokumen teks dengan uji *Kolmogorov-Smirnov* (K-S). Dimana penggunaan *term* pertama dalam setiap kalimat telah membuktikan munculnya perbedaan pola seseorang dalam berkomunikasi menggunakan teks.

Sedangkan jarak (*d*) antara munculnya pasangan *term* pertama dan frekuensi munculnya pasangan *term* dari dua dokumen dihitung untuk mengetahui kesamaan pola dokumen teks (Soehardjoepri dkk., 2015). Juga telah menguatkan ide dan pendugaan adanya perbedaan pola dokumen teks yang dibangun oleh cara komunikasi seseorang menggunakan teks.

Kedua hasil di atas perlu dikembangkan ke deteksi pola munculnya *term* dengan mempertimbangkan adanya faktor ketidakpastian suatu *term* muncul mengikuti *term-term* lainnya. Oleh sebab itu dalam bab ini dikaji urutan munculnya tiga *term* pertama dalam setiap kalimat dengan melibatkan faktor probabilistik munculnya *term* yang mewarnai pola suatu dokumen teks.

Dalam bab ini dibahas pendekatan lain dalam mengidentifikasi kesamaan pola dokumen teks dengan melihat munculnya tiga *term* pertama dari hasil *parsing text*. Hal ini dilakukan agar identifikasi pola dokumen teks lebih akurat dibandingkan dengan identifikasi pola dokumen sebelumnya yaitu identifikasi pola dokumen teks berdasarkan munculnya *term* pertama dari suatu dokumen teks dan munculnya pasangan *term* pertama dan frekuensinya dari dua dokumen teks.

Munculnya tiga *term* pertama pada setiap kalimat dalam masing-masing dokumen dibangun sebagai struktur BN dan menerapkan prinsip *likelihood* untuk mengukur kesamaan pola dokumen teks.

6.1 Munculnya Term

Latent Semantic Analysis (LSA) memiliki kontribusi yang signifikan dalam mendeteksi kesamaan dokumen. Kemampuan untuk mengekstrak dan mewakili makna kontekstual penggunaan kata dalam dokumen teks dapat diterapkan pada corpus besar (Landauer dkk., 1998). Hasil *parsing text* yang merupakan bagian dari LSA, berupa *term-term* di setiap kalimat dalam masing-masing dokumen dapat disusun dalam order statistik munculnya *term*. Order statistik munculnya *term* dilambangkan,

$$T_{(1)}, T_{(2)}, T_{(3)}, ..., T_{(n)}.$$

Dimana subskrip (i) dalam $T_{(i)}$ menunjukkan order statistik ke-i munculnya *term*. Order statistik pertama (order statistik terkecil) selalu merupakan minimum sampel *term*, yaitu,

$$T_{(1)} = \min \{T_1, ..., T_n\}.$$

Demikian pula, untuk sampel dengan ukuran n, order statistik ke-n (order statistik terbesar) adalah maksimum sampel *term*, yaitu,

$$T_{(n)} = \text{maks } \{T_1, ..., T_n\}.$$

Berdasar order statistik munculnya *term* dibentuklah algoritma tentang order munculnya *term* yang dirinci dalam Algoritma 6.1.

Algoritma 6.1: Order Munculnya *Term*

Langkah 1. Mengambil $\forall T_i \in L$, L adalah kamus *term*,

Langkah 2. Menempatkan *term-term* ke dalam order munculnya *term* $T_{(1)}$, $T_{(2)}, T_{(3)}, ..., T_{(n)}$ setiap kalimat di masing-masing dokumen teks,

Langkah 3. Menandai order munculnya *term* setiap kalimat untuk masing-masing dokumen dengan nomer urut.

Algoritma ini menghasilkan tabel order munculnya *term* disetiap kalimat (yang mengandung *term-term* hasil *parsing text*) dalam masing-masing dokumen yang terdapat dalam Tabel 6.1.

Tabel 6.1. Order munculnya *term*

	Kalimat	**Order munculnya *term* ke-**									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Dok-1	1	T_{19}	T_2	T_1							
	2	T_9	T_7	T_1	T_2	T_1	T_3				
	3	T_1	T_4	T_{18}	T_9	T_8	T_8	T_6	T_2		
	4	T_1	T_5	T_7	T_2	T_{18}	T_6	T_2	T_5	T_4	T_3
Dok-2	**Kalimat**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	1	T_1	T_4	T_{18}	T_9	T_8	T_8	T_6	T_2		
	2	T_1	T_5	T_7	T_2	T_{18}	T_6	T_2	T_5	T_4	T_3
	3	T_{19}	T_2	T_1							
	4	T_9	T_7	T_1	T_2	T_1	T_3				
Dok-3	**Kalimat**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	1	T_{17}	T_{12}	T_{10}	T_{16}	T_{15}	T_{11}	T_{10}			
	2	T_{10}	T_{15}	T_{12}							
	3	T_{10}	T_{10}	T_{19}							
	4	T_{10}	T_{11}	T_{11}							
	5	T_{13}	T_{14}	T_{11}	T_{10}	T_{16}					
	6	T_{13}	T_{13}	T_{14}	T_{12}	T_{12}					
Dok-4	**Kalimat**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	1	T_{10}	T_{17}	T_{12}	T_{17}	T_{16}	T_{15}	T_{11}	T_{10}		
	2	T_{12}	T_{10}	T_{15}							
	3	T_{10}	T_{19}	T_{10}							
	4	T_{10}	T_{11}	T_{11}							
	5	T_{11}	T_{10}	T_{16}	T_{14}	T_{13}					
	6	T_{12}	T_{14}	T_{13}	T_{12}	T_{13}					
Dok-5	**Kalimat**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	1	T_{10}	T_{17}	T_{12}	T_{17}	T_{16}	T_{15}	T_{11}	T_{10}		
	2	T_{12}	T_{10}	T_{15}							
	3	T_{10}	T_{19}	T_{10}							
	4	T_{10}	T_{11}	T_{11}							
	5	T_{11}	T_{10}	T_{16}	T_{14}	T_{13}					
	6	T_{12}	T_{14}	T_{13}	T_{12}	T_{13}					
	7	T_{19}	T_2	T_1							
	8	T_9	T_7	T_1	T_2	T_1	T_3				
	9	T_1	T_4	T_{18}	T_9	T_8	T_8	T_6	T_2		
	10	T_1	T_5	T_7	T_2	T_{18}	T_6	T_2	T_5	T_4	T_3

6.2 *Bayesian Network* (BN)

Order munculnya *term* yang dihasilkan dari Algoritma 6.1 didefinisikan sebagai struktur munculnya *term* dalam kalimat. Stuktur munculnya *term* ini, merupakan order *term* yang tersusun dalam *network* setiap kalimat dalam masing-masing dokumen. Dalam struktur kalimat, munculnya *term* yang kedua tergantung dari munculnya *term* kesatu, demikian pula untuk munculnya *term* berikutnya tergantung dari munculnya *term* sebelumnya. Munculnya *term* ini

didefinisikan sebagai simpul. Jaringan simpul mewakili order munculnya *term* dalam kalimat masing-masing dokumen merupakan peristiwa BN. Munculnya *term* pertama disetiap kalimat dapat dihitung probabilitasnya diantara munculnya *term-term* yang pertama. Munculnya *term* kedua dapat dihitung probabilitasnya diantara munculnya *term-term* yang kedua. Probabilitas munculnya *term* berikutnya dapat dihitung probabilitasnya dengan cara yang sama. Probabilitas munculnya *term* yang mewakili order munculnya *term* pertama sampai munculnya *term* terakhir di setiap kalimat dapat dihitung.

Mempertimbangkan BN yang mengandung n simpul, yaitu simpul X_1 ke X_n, diambil dari kamus *term*, maka terjadinya join antar simpul-simpul dapat di representasikan sebagai $P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$ atau $P(x_1, x_2, \dots, x_n)$. Order kemunculan *term* dalam BN menunjukkan order bersyarat munculnya *term* setiap kalimat dalam masing-masing dokumen. Konsep probabilitas bersyarat ini digunakan untuk menghitung probabilitas struktur BN. Aturan rantai teori probabilitas memungkinkan kita untuk menfaktorisasi probabilitas bersama menjadi,

$$P(x_1, x_2, ..., x_n) = P(x_1).P(x_2 \mid x_1). \dots .P(x_n \mid x_1, ..., x_{n-1})$$
$$= \prod_{i=1}^{n} P(x_i | x_1, \dots, x_{i-1}). \quad (6.1)$$

Order munculnya *term* dalam BN bisa mengikuti proses Markov, yaitu munculnya *term* berikutnya tergantung dari munculnya *term* sebelumnya. Oleh karena itu, struktur BN dalam persamaan (6.1) menyiratkan bahwa probabilitas bersyarat dari simpul tertentu tergantung hanya pada simpul induknya. Persamaan (6.1) dapat ditulis sebagai,

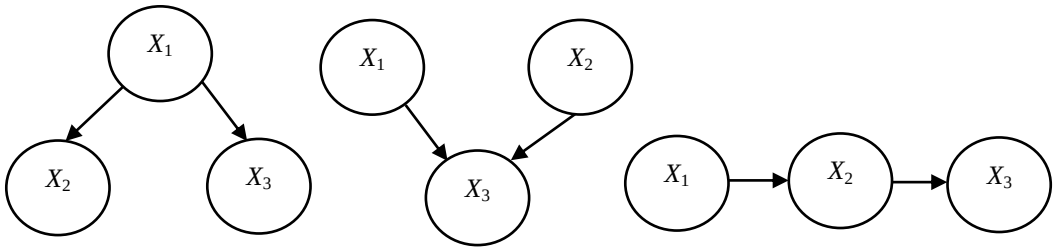
$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} P\big(x_i \big| Parents\ (X_i)\big). \quad (6.2)$$

dimana $Parents(X_i) \subseteq \{X_1, \dots, X_{i-1}\}$ (Kevin dan Ann, 2011).

Untuk n = 3 simpul, simpul yang dinyatakan dalam X_1, X_2, dan X_3, sehingga terdapat tiga kemungkinan struktur BN seperti yang ditunjukkan dalam *Directed Acyclic Graph* (DAG) pada Gambar 6.1. Distribusi bersama menurut DAG ini

akan mengikuti model yang diberikan dalam Chow dan Liu (1968), sebagai contoh pohon ketergantungan atau sebagai simpul yang terhubung DAG yang mewakili sifat Markov. Distribusi gabungan simpul X_1, X_2, dan X_3 dari setiap DAG pada Gambar 6.1 sebagai berikut:

1. Gambar 6.1.(a): $P(x_1, x_2, x_3) = P(x_2|x_1).P(x_3|x_1).P(x_1)$,
2. Gambar 6.1.(b): $P(x_1, x_2, x_3) = P(x_3|x_1, x_2).P(x_1).P(x_2)$,
3. Gambar 6.1.(c): $P(x_1, x_2, x_3) = P(x_3|x_2).P(x_2|x_1).P(x_1)$.



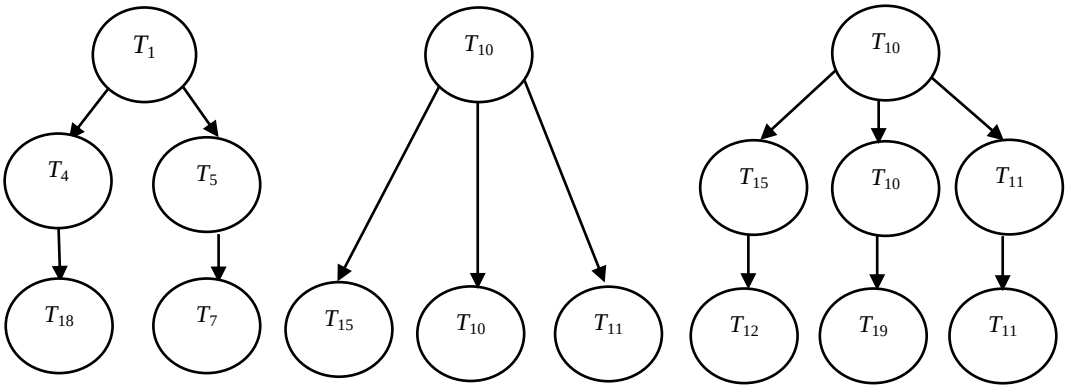
(a) (b) (c)

Gambar 6.1 Directed Acyclic Graph (DAG) untuk tiga simpul

Misalkan tiga kalimat yang digunakan untuk mengekspresikan munculnya *term* yang diambil dari kamus *term* yang ditunjukkan pada Gambar 6.2 dengan mengubah X_i pada Gambar 6.1 menjadi T_i sesuai dengan kemunculan *term* dalam kamus *term*. Distribusi gabungan *term* dari setiap grafik pada Gambar 6.2 dapat ditulis sebagai berikut:

1. Gambar 6.2.(a):
   $P(T_1, T_4, T_5, T_{18}, T_7) = [P(T_7|T_5).P(T_5|T_1)].[P(T_{18}|T_4).P(T_4|T_1)].P(T_1)$,
2. Gambar 6.2.(b): $P(T_{10}, T_{15}, T_{10}, T_{11}) = P(T_{15}|T_{10}).P(T_{10}|T_{10}).P(T_{11}|T_{10}).P(T_{10})$,
3. Gambar 6.2.(c):
   $P(T_{10}, T_{15}, T_{10}, T_{11}, T_{12}, T_{19}, T_{11})$
   $= (T_{12}|T_{15}) . P(T_{15}|T_{10}) . P(T_{19}|T_{10}) . P(T_{10}|T_{10}).P(T_{11}|T_{11}).P(T_{11}|T_{10}).P(T_{10})$

Setiap kalimat mengandung beberapa *term* yang tersusun secara berurutan. Dalam Bahasa Indonesia munculnya tiga *term* pertama umumnya mengandung pokok utama dari suatu kalimat yaitu subjek, predikat, dan objek. Dalam penelitian ini, munculnya tiga *term* pertama setiap kalimat yang diambil untuk diproses sebagai pola kalimat. Pembuatan algoritma tentang pola munculnya *term* dirinci dalam Algoritma 6.2.



(a) (b) (c)

Gambar 6.2 DAG untuk 3 kalimat

Setiap kalimat mengandung beberapa *term* yang tersusun secara berurutan. Dalam Bahasa Indonesia munculnya tiga *term* pertama umumnya mengandung pokok utama dari suatu kalimat yaitu subjek, predikat, dan objek. Dalam penelitian ini, munculnya tiga *term* pertama setiap kalimat yang diambil untuk diproses sebagai pola kalimat. Pembuatan algoritma tentang pola munculnya *term* dirinci dalam Algoritma 6.2.

Algoritma 6.2: Pola Munculnya *Term*

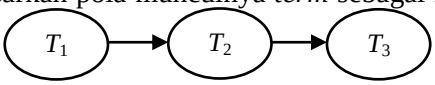
Langkah 1.	Mengambil semua *term* yang tepat di setiap kalimat dari order munculnya *term* $T_{(1)}$, $T_{(2)}$, $T_{(3)}$, ..., $T_{(n)}$,
Langkah 2.	Mengatur semua *term-term* mengikuti struktur setiap kalimat dalam setiap dokumen sebagai pola munculnya *term*.
Langkah 3.	Mengambil tiga *term* pertama (T_1,T_2,T_3) dari pola munculnya *term* di setiap kalimat dari setiap dokumen.
Langkah 4.	Membangun gabungan jaringan *term* dari pola munculnya tiga *term* pertama (T_1,T_2,T_3) dalam setiap kalimat dari setiap dokumen.

Munculnya *term* dalam kalimat selalu tergantung pada munculnya *term* sebelumnya. Rangkaian munculnya *term* akan membangun jaringan *term*. Jaringan *term* ini bisa disebut sebagai BN karena munculnya serangkaian *term* menunjukkan salah satu datang sebelum yang lainnya. Oleh karena itu, serangkaian *term* akan mewakili serangkaian dari urutan *prior* dan *posterior* yang merupakan BN. Algoritma 6.3 menunjukkan langkah-langkah untuk menghitung *likelihood* munculnya tiga *term* pertama sebagai distribusi gabungan dari *term*

diletakkan dalam kalimat sebagai *Bayesian Network*. Pembuatan algoritma tentang *Likelihood* munculnya tiga *term* pertama dirinci dalam Algoritma 6.3. Algoritma ini akan menjelaskan bagaimana menghitung distribusi probabilitas gabungan dari munculnya tiga *term* pertama dalam setiap kalimat sebagai *likelihood*nya.

Algoritma 6.3: *Likelihood* Munculnya Tiga *Term* Pertama

Langkah 1. Mengambil munculnya tiga *term* pertama (T_1,T_2,T_3) dari setiap kalimat berdasarkan pola munculnya *term* sebagai representasi dari kalimat,



Langkah 2. Mengumpulkan dan menghitung individu *term*

- Mengumpulkan dan menghitung frekuensi $f(T_i)$ munculnya *term* pertama dari $\forall k$, k = kalimat sebagai sebuah kelompok kemudian menghitung $p(T_{(i)})$, $\forall T_{(i)}$, $T_{(i)}$ adalah *term* antara munculnya *term* pertama dalam kelompok ini.
- Kumpulkan munculnya *term* kedua dan ketiga dari semua kalimat dan menghitung probabilitas $p(T_i)$, dalam kelompok mereka seperti yang dilakukan untuk *term* pertama.

Langkah 3. Menghitung *likelihood* setiap kalimat dengan mengalikan semua probabilitas munculnya tiga *term* pertama dalam kalimat.
$P(T_1, T_2, T_3) = P(T_3|T_2,T_1).P(T_2|T_1).P(T_1)$

Dokumen uji yang berisi beberapa kalimat, yang dapat dihitung *likelihood*nya berdasarkan hukum probabilitas dihitung dengan Algoritma 6.3. Selain itu, prinsip urutan kalimat dalam dokumen tersebut juga dapat diasumsikan sebagai struktur BN dengan kalimat diasumsikan independen. Probabilitas pola dokumen dapat diwakili oleh probabilitas urutan munculnya *term* di setiap kalimat dihitung dengan perkalian dari semua *likelihood* setiap kalimat. Ketika semua dokumen dapat dilihat *likelihood* dokumenya, maka dokumen tersebut dapat dibandingkan pola munculnya *term*nya dengan menghitung rasio *likelihood* (L_R) masing-masing pasangan dokumen. Algoritma untuk menghitung rasio *likelihood* antara dua dokumen dapat dilihat pada Algoritma 6.4.

Algoritma 6.4: Rasio *Likelihood* Dokumen

Langkah 1. Menghitung *likelihood* setiap kalimat diwakili oleh tiga *term* yang pertama dalam setiap dokumen, menggunakan Algoritma 6.3.
$P(T_1, T_2, T_3) = P(T_3|T_2,T_1).P(T_2|T_1).P(T_1)$

Langkah 2. Menghitung *likelihood* setiap dokumen dengan mengalikan *likelihood* setiap kalimat dalam dokumen terkait.

Langkah 3. Menghitung **harga mutlak** rasio *likelihood* ($|L_R|$) dua dokumen dengan membagi *likelihood* dokumen uji dengan dokumen yang lainnya.

Langkah 4. Menentukan perbandingan antara kedua dokumen dengan menggunakan rasio kemungkinan mereka berdasarkan *Bayes Factor* yang sesuai dengan Tabel 6.2.

Tabel 6.2. *Bayes Factor*

L_R	Hasil
1 to 3	Tidak layak dikatakan ada perbedaan
3 to 20	Positif adanya perbedaan
20 to 150	Kuat adanya perbedaan
>150	Sangat Kuat adanya perbedaan

Sumber: Kass dan Raftery, 1995

di mana L_R = rasio *likelihood* Dok-i dan Dok-j ($j \neq i$), untuk i = 1, ..., k dan k adalah jumlah dokumen.

Algoritma 6.4 akan terlebih dahulu diterapkan untuk menguji pola kesamaan lima dokumen yang diuji seperti yang telah digunakan di Soehardjoepri dkk. (2013) dan dilanjutkan dengan dokumen keenam yang berbeda dengan kelima dokumen sebelumnya.

6.3 Implementasi Numerik

Lima dokumen yang telah digunakan dalam Soehardjoepri dkk. (2013), digunakan lagi pada implementasi numerik untuk menunjukkan hasil Algoritma 6.4. Pertama, dari lima dokumen akan dikupas semua persyaratan mereka dengan menggunakan LSA sebagaimana tercantum dalam Algoritma 6.1. Hasil algoritma ini adalah munculnya *term* dalam setiap kalimat dari setiap dokumen, yang ditunjukkan pada Tabel 6.1.

Pelaksanaan Algoritma 6.2 dengan aturan berdasarkan munculnya tiga *term* pertama dalam kalimat Indonesia untuk kelima dokumen yang memiliki urutan

term tercantum dalam Tabel 6.1, memberikan struktur pola munculnya *term* seperti yang ditunjukkan pada kolom kedua dari Tabel 6.3 sampai dengan Tabel 6.7. Dengan menerapkan dalam bentuk BN, probabilitas struktur setiap kalimat yang diberikan pada kolom ketiga dari tabel tersebut.

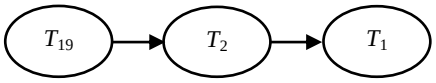
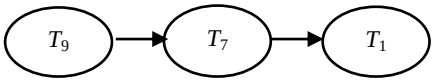
Dalam rangka untuk mencari probabilitas setiap struktur kalimat, Algoritma 6.3 dapat diterapkan. Perhitungan probabilitas setiap *term* dalam struktur harus disiapkan dengan menghitung frekuensi setiap munculnya *term* yang muncul dalam setiap kalimat dari lima dokumen secara keseluruhan. Frekuensi setiap munculnya *term* dalam setiap kalimat dari lima dokumen dapat dilihat pada Tabel 6.8. Berdasarkan frekuensi ini, probabilitas setiap kejadian *term* dapat ditemukan dengan menghitung *term* yang terkait dalam setiap kemunculan, seperti yang dinyatakan dalam langkah kedua dari Algoritma 6.3. Dengan menggunakan Tabel 6.8 dan memasukkan ke langkah terakhir dari Algoritma 6.3, probabilitas munculnya tiga *term* pertama untuk setiap kalimat dalam lima dokumen ditunjukkan pada Tabel 6.9.

Tabel 6.3. Struktur pola munculnya *term* dan probabilitas untuk Dok-1

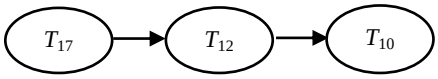
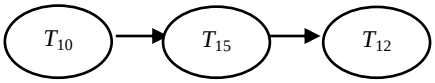
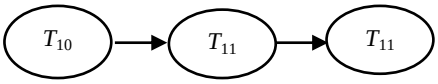
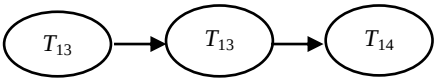
No.	Struktur pola munculnya *term*	Probabilitas gabungan dari struktur
1.	$T_{19} \rightarrow T_2 \rightarrow T_1$	$P(T_{19}, T_2, T_1) = P(T_1\|T_2,T_{19}).P(T_2\|T_{19}).P(T_{19})$
2.	$T_9 \rightarrow T_7 \rightarrow T_1$	$P(T_9, T_7, T_1) = P(T_1\|T_7, T_9).P(T_7\|T_9).P(T_9)$
3.	$T_1 \rightarrow T_4 \rightarrow T_{18}$	$P(T_1, T_4, T_{18}) = P(T_{18}\|T_4, T_1).P(T_4\|T_1).P(T_1)$
4.	$T_1 \rightarrow T_5 \rightarrow T_7$	$P(T_1, T_5, T_7) = P(T_7\|T_5, T_1).P(T_5\|T_1).P(T_1)$

Tabel 6.4. Struktur pola munculnya *term* dan probabilitas untuk Dok-2

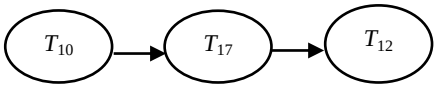
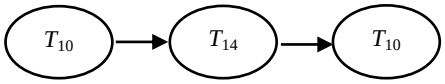
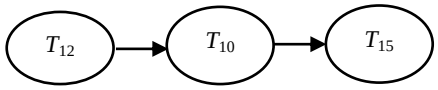
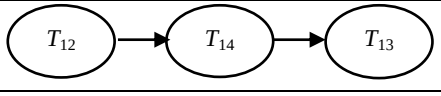
No.	Struktur pola munculnya *term*	Probabilitas gabungan dari struktur
1.	$T_1 \rightarrow T_4 \rightarrow T_{18}$	$P(T_1, T_4, T_{18}) = P(T_{18}\|T_4, T_1).P(T_4\|T_1).P(T_1)$
2.	$T_1 \rightarrow T_5 \rightarrow T_7$	$P(T_1, T_5, T_7) = P(T_7\|T_5, T_1).P(T_5\|T_1).P(T_1)$

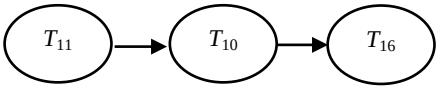
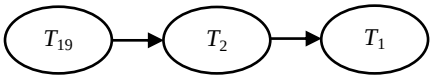
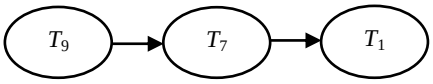
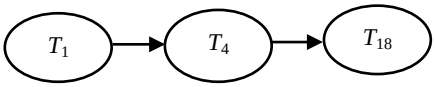
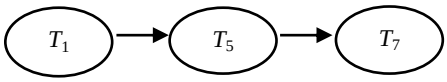
No.	Struktur pola munculnya *term*	Probabilitas gabungan dari struktur
3.		$P(T_{19}, T_2, T_1) = P(T_1\|T_2,T_{19}).P(T_2\|T_{19}).P(T_{19})$
4.		$P(T_9, T_7, T_1) = P(T_1\|T_7,T_9).P(T_7\|T_9).P(T_9)$

Tabel 6.5. Struktur pola munculnya *term* dan probabilitas untuk Dok-3

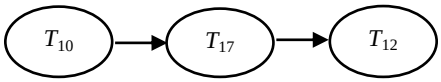
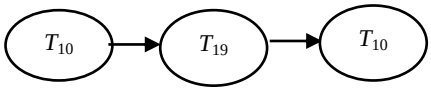
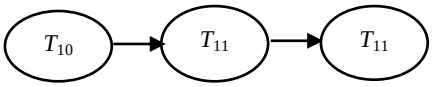
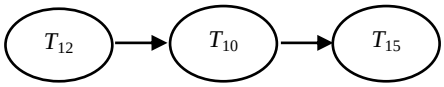
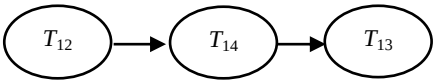
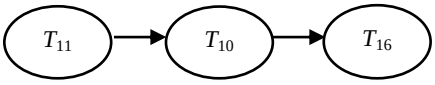
No.	Struktur pola munculnya *term*	Probabilitas gabungan dari struktur
1.		$P(T_{17}, T_{12}, T_{10}) = P(T_{10}\|T_{12},T_{17}).P(T_{12}\|T_{17}).P(T_{17})$
2.		$P(T_{10}, T_{15}, T_{12}) = P(T_{12}\|T_{15},T_{10}).P(T_{15}\|T_{10}).P(T_{10})$
3.		$P(T_{10}, T_{10}, T_{19}) = P(T_{19}\|T_{10},T_{10}).P(T_{10}\|T_{10}).P(T_{10})$
4.		$P(T_{10}, T_{11}, T_{11}) = P(T_{11}\|T_{11},T_{10}).P(T_{11}\|T_{10}).P(T_{10})$
5.		$P(T_{13}, T_{14}, T_{11}) = P(T_{11}\|T_{14},T_{13}).P(T_{14}\|T_{13}).P(T_{13})$
6.		$P(T_{13}, T_{13}, T_{14}) = P(T_{14}\|T_{13},T_{13}).P(T_{13}\|T_{13}).P(T_{13})$

Tabel 6.6. Struktur pola munculnya *term* dan probabilitas untuk Dok-5

No.	Struktur pola munculnya *term*	Probabilitas gabungan dari struktur
1.		$P(T_{10}, T_{17}, T_{12}) = P(T_{12}\|T_{17},T_{10}).P(T_{17}\|T_{10}).P(T_{10})$
2.		$P(T_{10}, T_{14}, T_{10}) = P(T_{10}\|T_{14},T_{10}).P(T_{14}\|T_{10}).P(T_{10})$
3.		$P(T_{10}, T_{11}, T_{11}) = P(T_{11}\|T_{11},T_{10}).P(T_{11}\|T_{10}).P(T_{10})$
4.		$P(T_{12}, T_{10}, T_{15}) = P(T_{15}\|T_{10},T_{12}).P(T_{10}\|T_{12}).P(T_{12})$
5.		$P(T_{12}, T_{14}, T_{13}) = P(T_{13}\|T_{14},T_{12}).P(T_{14}\|T_{12}).P(T_{12})$

No.	Struktur pola munculnya *term*	Probabilitas gabungan dari struktur
6.		$P(T_{11}, T_{10}, T_{16}) = P(T_{16}\|T_{10},T_{11}).P(T_{10}\|T_{11}).P(T_{11})$
7.		$P(T_{19}, T_2, T_1) = P(T_1\|T_2,T_{19}).P(T_2\|T_{19}).P(T_{19})$
8.		$P(T_9, T_7, T_1) = P(T_1\|T_7,T_9).P(T_7\|T_9).P(T_9)$
9.		$P(T_1, T_4, T_{18}) = P(T_{18}\|T_4,T_1).P(T_4\|T_1).P(T_1)$
10.		$P(T_1, T_5, T_7) = P(T_7\|T_5,T_1).P(T_5\|T_1).P(T_1)$

Tabel 6.7. Struktur pola munculnya *term* dan probabilitas untuk Dok-4

No.	Struktur pola munculnya *term*	Probabilitas gabungan dari struktur
1.		$P(T_{10}, T_{17}, T_{12}) = P(T_{12}\|T_{17},T_{10}).P(T_{17}\|T_{10}).P(T_{10})$
2.		$P(T_{10}, T_{19}, T_{10}) = P(T_{10}\|T_{19},T_{10}).P(T_{19}\|T_{10}).P(T_{10})$
3.		$P(T_{10}, T_{11}, T_{11}) = P(T_{11}\|T_{11},T_{10}).P(T_{11}\|T_{10}).P(T_{10})$
4.		$P(T_{12}, T_{10}, T_{15}) = P(T_{15}\|T_{10},T_{12}).P(T_{10}\|T_{12}).P(T_{12})$
5.		$P(T_{12}, T_{14}, T_{13}) = P(T_{13}\|T_{14},T_{12}).P(T_{14}\|T_{12}).P(T_{12})$
6.		$P(T_{11}, T_{10}, T_{16}) = P(T_{16}\|T_{10},T_{11}).P(T_{10}\|T_{11}).P(T_{11})$

Adapun banyaknya frekuensi munculnya term pertama, kedua dan ketiga untuk setiap kalimat di 5 dokumen terdapat dalam Tabel 6.8.

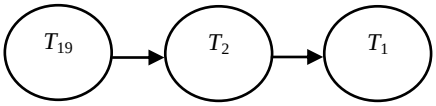
Tabel 6.8. Frekuensi munculnya *term* dari lima dokumen

Urutan terjadinya					
ke-1		ke-2		ke-3	
Term	Freq	*Term*	Freq	*Term*	Freq
T_{10}	9	T_{10}	5	T_1	6
T_1	6	T_7	3	T_{11}	4
T_{12}	4	T_4	3	T_{18}	3
T_9	3	T_5	3	T_7	3
T_{19}	3	T_{11}	3	T_{10}	3
T_{11}	2	T_{14}	3	T_{12}	3
T_{13}	2	T_2	3	T_{15}	2
T_{17}	1	T_{12}	1	T_{16}	2
		T_{17}	2	T_{13}	2
		T_{19}	2	T_{19}	1
		T_{15}	1	T_{14}	1
		T_{13}	1		

Perhitungan *posterior* berdasarkan Teorema Bayes untuk parameter $\boldsymbol{\theta}$ (persamaan 2.7) diberikan:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{y})}$$

Sehingga sebuah kalimat yang tersusun atas beberapa order *term* akan tersusun pula serangkaian urutan *prior* dan *posterior*. Misalkan dalam Tabel 6.3 untuk kalimat 1 dalam Dok-1, dengan order *term* sebagai berikut:



Kejadian T_{19} dengan probabilitas $P(T_{19})$ akan menjadi *prior* terjadinya T_2 dengan probabilitas $P(T_2)$. Sehingga probabilitas $P(T_2)$ menunjukan *posterior* dan $P(T_{19})$ sebagai *prior*nya. Selanjutnya $P(T_2)$ akan berfungsi sebagai *prior* dalam proses terjadinya T_1 sebesar $P(T_1)$. Dengan cara yang sama, suatu order *term* dalam sebuah kalimat dapat dihitung *prior* dan *posterior* yang lainnya. *Likelihood* suatu kalimat terbentuknya adalah perkalian probabilitas munculnya *term* pertama, kedua dan ketiga, sehingga *likelihood* untuk kalimat 1 dalam Dok-1 adalah $P(T_{19}, T_2, T_1) = P(T_1|T_2, T_{19}).P(T_2|T_{19}).P(T_{19})$ = 0,00100.

Perhitungan *likelihood* setiap kalimat dalam Dok-1 sd Dok-5 terdapat dalam Tabel 6.9. Membandingkan antara dua dokumen, misalnya Dok-1 dan Dok-2 yang dirancang sebelumnya untuk memiliki struktur yang sama, dapat dilakukan dengan menghitung rasio *likelihood* mereka. *Likelihood* setiap dokumen dapat dihitung dengan menggunakan Tabel 6.9 dan memasukkan ke Algoritma 6.4. Andaikan Dok-1 mempunyai pola yang berbeda dengan Dok-2, rasio *likelihood* dapat ditemukan dengan membagi likelihood Dok-1 dengan likelihood Dok-2. Berdasarkan Tabel 6.9, *likelihood* Dok-1 dan Dok-2 adalah masing-masing $4{,}00 \times 10^{-12}$, maka rasio *likelihood* adalah 1 ($L_R = 1$) dapat dilihat pada Tabel 6.12 dan hipotesisnya H_0 diterima (yang mempunyai kesimpulan bahwa Dok-1 dan Dok-2 memiliki pola kesamaan munculnya tiga *term* pertama sama).

Tabel 6.9. *Likelihood* setiap kalimat dalam setiap dokumen

Dokumen	Kalimat	*Likelihood*
Dok-1	1	0,00100
	2	0,00100
	3	0,00200
	4	0,00200
Dok-2	1	0,00200
	2	0,00200
	3	0,00100
	4	0,00100
Dok-3	1	0,00004
	2	0,00033
	3	0,00033
	4	0,00300
	5	0,00007
	6	0,00007

Dokumen	Kalimat	*Likelihood*
Dok-4	1	0,00133
	2	0,00059
	3	0,00133
	4	0,00300
	5	0,00030
	6	0,00059
Dok-5	1	0,00133
	2	0,00059
	3	0,00133
	4	0,00300
	5	0,00030
	6	0,00059
	7	0,00011
	8	0,00100
	9	0,00200
	10	0,00200

Perbandingan antara pasangan dokumen-dokumen yang lainnya diperoleh dengan cara yang sama, dan rasio *likelihood* dapat dilihat pada Tabel 6.12. Jika L_R semakin menjauh dari 1 ($L_R > 1$), maka kedua dokumen semakin ada perbedaan. Rasio *likelihood* lain menunjukkan bahwa H_0 ditolak dan dokumen-dokumen memiliki pola yang berbeda secara signifikan (Chow dan Liu, 1968).

Tabel 6.10. Probabilitas munculnya *term* untuk dokumen (Penyebut)

SEBAGAI PEMBILANG			
Dok-1	Dok-2	Dok-3	Dok-4
$4{,}00x10^{-12}$	$4{,}00x10^{-12}$	$6{,}77x10^{-23}$	$5{,}55x10^{-19}$

Tabel 6.11. Probabilitas munculnya *term* untuk dokumen (Pembilang)

SEBAGAI PENYEBUT				
Dok-1	Dok-2	Dok-3	Dok-4	Dok-5
$4{,}00x10^{-12}$	$4{,}00x10^{-12}$	$6{,}77x10^{-23}$	$5{,}55x10^{-19}$	$2{,}47x10^{-31}$

Tabel 6.12. Rasio *Likelihood* (L_R) untuk pasangan dokumen dari 5 dokumen

		Pembilang				
	L_R	Dok-1	Dok-2	Dok-3	Dok-4	Dok-5
Penyebut	Dok-1		1,00	$5{,}90x10^{+10}$	$7{,}21x10^{+06}$	$1{,}62x10^{+19}$
	Dok-2			$5{,}90x10^{+10}$	$7{,}21x10^{+06}$	$1{,}62x10^{+19}$
	Dok-3				8192,00	$2{,}75x10^{+08}$
	Dok-4					$2{,}25x10^{+12}$

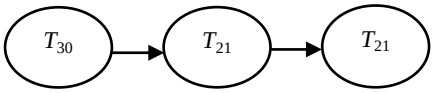
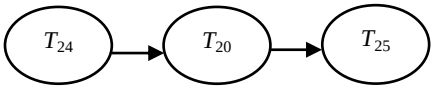
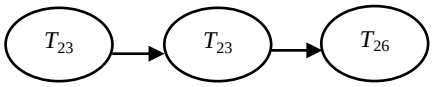
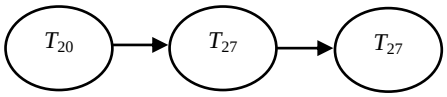
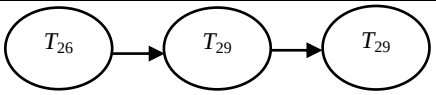
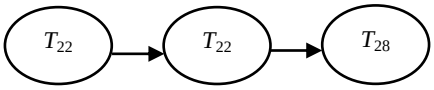
Untuk menunjukkan hasil metode yang diusulkan ini, kita mengambil dokumen lain (Dok-6) yang dirancang sebagai hampir mirip dengan Dok-5 tetapi dirancang sebagai benar-benar berbeda dengan keempat dokumen sebelumnya. Setelah proses *parsing text* berhasil memproses keenam dokumen dan menerapkan Algoritma 6.1 diikuti dengan memotong munculnya tiga *term* dari mereka, maka pola munculnya *term* Dok-6 dapat dilihat pada Tabel 6.13.

Frekuensi setiap munculnya *term* dalam setiap kalimat dari enam dokumen dapat dilihat pada Tabel 6.14, merevisi informasi pada Tabel 6.8. Berdasarkan frekuensi pada Tabel 6.14 dan menerapkan Algoritma 6.3, *likelihood* setiap kalimat dalam setiap dokumen untuk semua enam dokumen akan berubah dari Tabel 6.9 menjadi Tabel 6.15. Pasangan rasio *likelihood* antara enam dokumen, dapat ditemukan dengan menerapkan Algoritma 4 dan hasilnya dapat dilihat pada Tabel 6.18.

Berdasarkan Tabel 6.18, pengujian pola kesamaan Dok-6 untuk lima dokumen sebelumnya tidak akan mengubah keputusan. Tabel ini juga menunjukkan bahwa Dok-6 secara signifikan berbeda dengan semua dokumen kecuali antara Dok-5 dan Dok-6. Kedua dokumen terakhir memiliki pola yang

hampir sama, karena rasio *likelihood* mereka yang dapat dinyatakan sebagai *'Not worth more than a bare mention'* seperti pada Tabel 6.2.

Tabel 6.13. Struktur pola munculnya *term* dan probabilitas untuk Dok-6

No.	Struktur pola munculnya *term*	Probabilitas gabungan dari struktur
1.		$P(T_{30},T_{21},T_{21}) = P(T_{21}\|T_{21},T_{30}).P(T_{21}\|T_{30}).P(T_{30})$
2.		$P(T_{24},T_{20},T_{25}) = P(T_{25}\|T_{20},T_{24}).P(T_{20}\|T_{24}).P(T_{24})$
3.		$P(T_{23},T_{23},T_{26}) = P(T_{26}\|T_{23},T_{23}).P(T_{23}\|T_{23}).P(T_{23})$
4.		$P(T_{20},T_{27},T_{27}) = P(T_{27}\|T_{27},T_{20}).P(T_{27}\|T_{20}).P(T_{20})$
5.		$P(T_{26},T_{29},T_{29}) = P(T_{29}\|T_{29},T_{26}).P(T_{29}\|T_{26}).P(T_{26})$
6.		$P(T_{22},T_{22},T_{28}) = P(T_{28}\|T_{22},T_{22}).P(T_{22}\|T_{22}).P(T_{22})$

Tabel 6.14. Frekuensi dari munculnya *term* dalam enam dokumen

Urutan terjadinya					
ke-1		ke-2		ke-3	
Term	Freq	*Term*	Freq	*Term*	Freq
T_{10}	9	T_{10}	5	T_1	6
T_1	6	T_7	3	T_{11}	4
T_{12}	4	T_4	3	T_{18}	3
T_9	3	T_5	3	T_7	3
T_{19}	3	T_{11}	3	T_{10}	3
T_{11}	2	T_{14}	3	T_{12}	3
T_{13}	2	T_2	2	T_{15}	2
T_{17}	1	T_{12}	2	T_{16}	2
T_{20}	1	T_{17}	2	T_{13}	2
T_{22}	1	T_{19}	2	T_{19}	1
T_{24}	1	T_{15}	1	T_{14}	1
T_{26}	1	T_{13}	1	T_{21}	1
T_{30}	1	T_{20}	1	T_{26}	1
T_{23}	1	T_{21}	1	T_{25}	1
		T_{22}	1	T_{27}	1
		T_{23}	1	T_{28}	1
		T_{27}	1	T_{29}	1
		T_{29}	1		

Table 6.15. *Likelihood* dari setiap kalimat dalam enam dokumen

Dokumen	Kalimat	*Likelihood*
Dok-1	1	0,00058
	2	0,00058
	3	0,00116
	4	0,00116
Dok-2	1	0,00116
	2	0,00116
	3	0,00058
	4	0,00058
Dok-3	1	0,00002
	2	0,00019
	3	0,00019
	4	0,00116
	5	0,00004
	6	0,00004
Dok-4	1	0,00077
	2	0,00034
	3	0,00077
	4	0,00116
	5	0,00017
	6	0,00034

Dokumen	Kalimat	*Likelihood*
Dok-5	1	0,00077
	2	0,00034
	3	0,00077
	4	0,00174
	5	0,00017
	6	0,00034
	7	0,00058
	8	0,00058
	9	0,00116
	10	0,00039
Dok-6	1	0,00002
	2	0,00002
	3	0,00002
	4	0,00002
	5	0,00002
	6	0,00002

Tabel 6.16. Probabilitas munculnya *term* untuk dokumen (Pembilang)

SEBAGAI PEMBILANG				
Dok-1	Dok-2	Dok-3	Dok-4	Dok-5
$4{,}49\text{x}10^{-13}$	$4{,}49\text{x}10^{-13}$	$1{,}70\text{x}10^{-24}$	$1{,}39\text{x}10^{-20}$	$3{,}12\text{x}10^{-33}$

Tabel 6.17. Probabilitas munculnya *term* untuk dokumen (Penyebut)

SEBAGAI PENYEBUT					
Dok-1	Dok-2	Dok-3	Dok-4	Dok-5	Dok-6
$4{,}49\text{x}10^{-13}$	$4{,}49\text{x}10^{-13}$	$1{,}70\text{x}10^{-24}$	$1{,}39\text{x}10^{-20}$	$3{,}12\text{x}10^{-33}$	$9{,}70\text{x}10^{-29}$

Dari Tabel 6.2 dan Tabel 6.18 menunjukkan Dok-1 dan Dok-2 nilai $L_R = 1$ yang mempunyai arti bahwa kedua dokumen ada kesamaan pola dokumen (*Not worth more than a bare mention*). Jika $L_R > 150$, maka kedua dokumen ada perbedaan pola dokumen. Perhitungan Rasio *Likelihood* antar dokumen telah dibuatkan sebuah program dengan panduan program terdapat dalam lampiran 5.

Tabel 6.18. Rasio *Likelihood* (L_R) untuk setiap pasangan dokumen dari enam dokumen

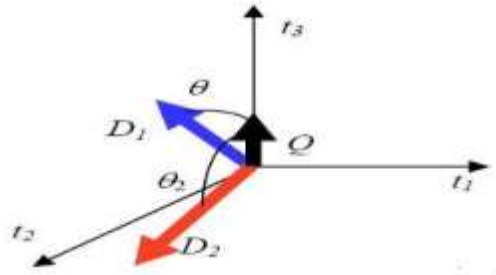
		Pembilang					
Penyebut	L_R	Dok-1	Dok-2	Dok-3	Dok-4	Dok-5	Dok-6
	Dok-1		1,00	$2{,}64x10^{+11}$	$3{,}23x10^{+07}$	$1{,}44x10^{+20}$	$4{,}63x10^{+15}$
	Dok-2			$2{,}64x10^{+11}$	$3{,}23x10^{+07}$	$1{,}44x10^{+20}$	$4{,}63x10^{+15}$
	Dok-3				8192,000	$5{,}44x10^{+08}$	$1{,}75x10^{+04}$
	Dok-4					$4{,}46x10^{+12}$	$1{,}43x10^{+08}$
	Dok-5						$3{,}11x10^{+04}$

Identifikasi dan kesamaan dokumen teks dengan memperhatikan pola struktur *term* (munculnya tiga *term* pertama) dari dua dokumen teks dapat ditentukan dengan menggunakan pendekatan *Bayesian Network* (BN) dan *Likelihood Ratio Test*. Hasil yang diperoleh 100 % sesuai dengan skenario dokumen asli.

BAB 7
PERBANDINGAN DETEKSI KESAMAAN POLA DOKUMEN TEKS DENGAN MODEL RUANG VEKTOR

Dalam bab ini disajikan perbandingan deteksi kesamaan Pola Dokumen teks dengan Model Ruang Vektor yang sudah diteliti oleh Sentosa tahun 2015. Ruang vektor adalah struktur matematika yang dibentuk oleh sekumpulan vektor, yaitu objek yang dapat dijumlahkan dan dikalikan dengan suatu bilangan, yang dinamakan bilangan skalar (Sentosa , 2015). Contoh ruang vektor adalah Vektor *Euclides* yang sering digunakan untuk melambangkan besaran fisika seperti gaya. Vektor-vektor yang berada di ruang R^n dikenal sebagai vektor *Euclides* sedangkan ruang vektornya disebut ruang n- *Euclides*. Model ruang vektor merupakan teknik dasar dalam perolehan informasi yang dapat digunakan untuk penelitian relevansi dokumen terhadap kata kunci pencarian (*query*) pada mesin pencarian, klasifikasi dokumen dan pengelompokan dokumen, sistem Temu-balik informasi (*information Retrieval System*).

Pencarian dalam sistem temu balik merupakan hal yang dibutuhkan, hal ini dikarenakan ketepatan pencarian sesuai *keyword* yang dimasukkan *user* dengan dokumen yang jumlahnya banyak. Model ruang vektor adalah suatu model yang digunakan untuk mengukur kemiripan antara suatu dokumen dengan suatu query. Query dan dokumen dianggap sebagai vektor-vektor pada ruang n-dimensi, dimana *t* adalah jumlah dari seluruh *term* yang ada dalam leksikon. Leksikon dalam penelitian ini dinamakan kamus *term* yang terindeks. Representasi vektor dalam ruang digambarkan di Gambar 7.1.



Gambar 7.1 Representasi vektor dalam ruang
(Sumber: Liyantanto, 2011)

Keterangan Gambar 7.1:

t_1, t_2, t_3	:	*Term-term* yang ada dalam kamus *term*.
θ	:	Sudut yang diapit oleh vektor $\boldsymbol{Q}$ dan vektor $\boldsymbol{D}_1$.
θ_2	:	Sudut yang diapit oleh vektor $\boldsymbol{Q}$ dan vektor $\boldsymbol{D}_2$.
$\boldsymbol{D}_1, \boldsymbol{D}_2$	:	Vektor dokumen uji.
$\boldsymbol{Q}$	:	Vektor dokumen query.

Selanjutnya akan dihitung *cosinus* sudut dari dua vektor, untuk digunakan mengukur kesamaan antara dua buah vektor yang dikenal dengan *Cosine Similiraty. Cosine Similarity* merupakan hasil *cosinus* dari sudut di antara kedua vektor. *Cosine Similarity* dapat dirumuskan sebagai berikut: (Munir, 2015)

$$S(Q,D) = \cos\theta = \frac{\boldsymbol{Q}.\boldsymbol{D}}{\|Q\|\|D\|} = \frac{\sum_{i=1}^{n} Q_i D_i}{\sqrt{\sum_{i=1}^{n} Q_i^2}\sqrt{\sum_{i=1}^{n} D_i^2}} \quad (7.1)$$

dimana,

$\boldsymbol{Q}$	:	Vektor *query* dokumen yang berisikan *n* elemen, dengan isi elemen-elemen vektornya adalah term-term dari query dokumen.
$\boldsymbol{D}$	:	Vektor dokumen uji yang berisikan *n* elemen, dengan isi elemen-elemen vektornya adalah term-term dari dokumen uji.
θ	:	Sudut yang diapit oleh vektor $\boldsymbol{Q}$ dan vektor $\boldsymbol{D}$.
$\boldsymbol{Q}.\boldsymbol{D}$	:	Hasil perkalian titik dari vektor $\boldsymbol{Q}$ dan vektor $\boldsymbol{D}$.
$\|Q\|$ dan $\|D\|$	:	Masing-masing adalah *Euclidean* dari vektor $\boldsymbol{Q}$ dan vektor $\boldsymbol{D}$. Panjang *Euclidean* diperoleh dari akar penjumlahan kuadrat elemen-elemen vektor tersebut.
$\cos\theta$	:	Nilai $\cos\theta$ antara 0 sd 1 ($0 \leq \cos\theta \leq 1$).

Cosine Similarity bernilai antara 0 sd 1. Jika nilainya 1, maka kedua dokumen adalah 100% sama. Sebaliknya jika nilainya 0, maka kedua dokumen adalah sangat berbeda.

Ambil Dok-1 dan Dok-2 dari 6 dokumen uji untuk dibandingkan kesamaannya dengan $\boldsymbol{Q}$ = vektor Dok-1 dan $\boldsymbol{D}_2$ = vektor Dok-2. Vektor $\boldsymbol{Q}$ berisikan *term-term* yang berada di Dok-1 (*D*1) dan *D*2 berisikan *term-term* yang berada di Dok-2. Frekuensi *term* untuk Dok-1 dan Dok-2 terdapat dalam Tabel 7.1.

Tabel 7.1. Frekuensi *term* Dok-1 dan Dok-2

Kode *Term*	*Term*	Frekuensi *term*	
		Dok-1	Dok-2
T_1	ruang	5	5
T_2	bunyi	5	5
T_3	gema	2	2
T_4	cegah	2	2
T_5	gedung	2	2
T_6	serap	2	2
T_7	keras	2	2
T_8	langit	2	2
T_9	dinding	2	2
T_{18}	bahan	2	2
T_{19}	akustik	1	1

Frekuensi *term* untuk Dok-1 dan Dok-2 direpresentasikan dalam vektor *D*1 dan vektor *D*2 dengan ukuran yang sama sebagai berikut:

$$D1 = \begin{pmatrix} 5 \\ 5 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 1 \end{pmatrix}, D2 = \begin{pmatrix} 5 \\ 5 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 1 \end{pmatrix}$$

Nilai pembilang dari *Cosine Similarity* adalah,

$D1.D2$ = (5)(5)+(5)(5)+(2)(2)+ (2)(2)+ (2)(2)+ (2)(2)+ (2)(2)+ (2)(2)+ (2)(2)+ (2)(2)+ (1)(1)

= 25+25+4+4+4+4+4+4+4+4+1

= 83

Nilai penyebut dari *Cosine Similarity* adalah,

$\|D1\| = \sqrt{5^2 + 5^2 + 2^2 + 2^2 + 2^2 + 2^2 + 2^2 + 2^2 + 2^2 + 2^2 + 1^2}$

= 9,110434

$\|D2\| = \sqrt{5^2 + 5^2 + 2^2 + 2^2 + 2^2 + 2^2 + 2^2 + 2^2 + 2^2 + 2^2 + 1^2}$

= 9,110434

Sehingga didapatkan persamaan (7.1) sebagai berikut:

$$\cos\theta = \frac{D1.D2}{\|D1\|\|D2\|}$$

$$= 83/(9{,}110434).(9{,}110434)$$

$$= 1$$

Nilai $\cos\theta$ diperoleh sebesar 1 yang berarti bahwa kedua dokumen (Dok-1 dan Dok-2) 100% sama.

Contoh berikutnya ambil Dok-1 dan Dok-3 yang terdapat dalam Tabel 7.2.

Tabel 7.2. Frekuensi *term* Dok-1 dan Dok-3

Kode *Term*	*Term*	Frekuensi *term*	
		Dok-1	Dok-3
T_{10}	musik	0	7
T_{11}	budaya	0	4
T_{12}	lampung	0	4
T_{13}	festival	0	3
T_{14}	ada	0	2
T_{15}	hingga	0	2
T_{16}	tradisional	0	2
T_{17}	daerah	0	1
T_{18}	bahan	2	0
T_{19}	akustik	1	1

Frekuensi *term* untuk Dok-1 dan Dok-3 direpresentasikan dalam vektor *D*1 dan vektor *D*3 dengan ukuran yang sama sebagai berikut:

$$D_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 2 \\ 1 \end{pmatrix}, D_3 = \begin{pmatrix} 7 \\ 4 \\ 4 \\ 3 \\ 2 \\ 2 \\ 2 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

Nilai pembilang dari *Cosine Similarity* adalah,

*D*1.*D*3 = (0)(7)+(0)(4)+(0)(4)+ (0)(3)+ (0)(2)+ (0)(2)+ (0)(2)+ (0)(1)+ (2)(0)+ (1)(1)

$D1.D3$ = 0+0+0+0+0+0+0+0+0+1

= 1

Nilai penyebut dari *Cosine Similarity* adalah,

$\|D1\| = \sqrt{0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 2^2 + 1^2}$

= 2,2361

$\|D3\| = \sqrt{7^2 + 4^2 + 4^2 + 3^2 + 2^2 + 2^2 + 2^2 + 1^2 + 0^2 + 1^2}$

= 10,1488

Sehingga didapatkan persamaan (7.1) sebagai berikut:

$$\cos\theta = \frac{D1.D3}{\|D1\|\|D3\|}$$

= 1/(2,2361).(10,1488)

= 0,04

Nilai $\cos\theta$ diperoleh sebesar 0,04 yang berarti bahwa kedua dokumen (Dok-1 dan Dok-3) hanya 4% yang sama. Dengan perhitungan yang sama untuk 6 dokumen uji didapatkan nilai $\cos\theta$ yang terdapat dalam Tabel 7.3. (Lampiran 4)

Tabel 7.3. Nilai $\cos\theta$ untuk pasangan dokumen uji

	D1	D2	D3	D4	D5	D6
D1		1	0,04	0,16	0,66	Δ
D2			0,04	0,16	0,66	Δ
D3				0,99	0,75	Δ
D4					0,75	Δ
D5						0

Keterangan:
Δ : tidak bisa diambil keputusan, karena penyebut dari $\cos\theta$ = 0

Nilai $\cos\theta$ dalam Tabel 7.3, juga bisa dinyatakan dalam prosentase yang terdapat dalam Tabel 7.4.

Tabel 7.4. Nilai $\cos\theta$ untuk pasangan dokumen uji (dalam %)

	D1	D2	D3	D4	D5	D6
D1		100%	4%	16%	66%	Δ
D2			4%	16%	66%	Δ
D3				99%	75%	Δ
D4					75%	Δ
D5						0%

Dari hasil Tabel 7.4 diperoleh bahwa Dok-1 dan Dok-2 nilai $\cos\theta$ = 100%, yang berarti kedua dokumen sama 100%. Sedang Dok-3 dan Dok-4 nilai $\cos\theta$ = 99% berarti kedua dokumen mempunyai kesamaan 99%. Nilai prosentase dari $\cos\theta$ merupakan nilai prosentase kesamaan kedua dokumen.

Cosine Similarity mempunyai kelemahan jika kedua dokumen sangat berbeda yang menyebabkan nilai penyebut $\cos\theta$ = 0, maka *Cosine Similarity* tidak bisa mengambil keputusan apakah kedua dokumen berbeda atau sama. Sedangkan hasil identifikasi kesamaan pola dokumen teks berdasarkan kemunculan *term* dalam kalimat bisa mengambil keputusan apakah kedua dokumen mempunyai kesamaan atau mempunyai perbedaan.

BAB 8
KESIMPULAN DAN SARAN

Dalam bab ini disajikan beberapa kesimpulan yang diambil dari uraian dalam bab-bab sebelumnya. Berdasarkan kesimpulan yang diperoleh tersebut, dapat dikemukakan beberapa saran bagi pengguna algoritma dan teorema yang dikembangkan dalam penelitian ini. Penelitian ini juga memiliki peluang pengkajian lebih lanjut untuk penelitian lain yang terkait dengan topik penelitian ini.

8.1 Kesimpulan

Dari pembahasan dalam Bab 4, Bab 5, dan Bab 6 dapat disimpulkan bahwa:

1. Perpaduan antara Algoritma 4.1, Algoritma 4.2 dan Algoritma 4.3 dengan uji pembeda K-S terbukti dapat digunakan untuk mengidentifikasi dan menguji kesamaan pola dokumen teks dengan memperhatikan munculnya *term* pertama di setiap kalimat dalam dokumen teks yang dibandingkan. Hasil perhitungan uji *Kolmogorov-Smirnov* (uji K-S) untuk 15 pasangan dokumen yang diuji terdapat 10 pasangan dokumen (66,67 %) sesuai dengan skenario enam dokumen asli.
2. Pengkombinasian antara Algoritma 5.1, Algoritma 5.2, dan Algoritma 5.3 dengan penghitungan jarak *Euclidean* dapat digunakan untuk mengidentifikasi dan menguji kesamaan pola dokumen teks dengan memperhatikan munculnya pasangan *term* pertama di setiap kalimat dalam dokumen teks yang dibandingkan. Hasil penghitungan jarak *Euclidean* didapatkan hanya 2 pasangan dokumen (Dok-3 dan Dok-4) yang tidak sesuai dengan skenario dokumen asli (6,67 %).
3. Penerapan Algoritma 6.1, Algoritma 6.2, Algoritma 6.3, dan Algoritma 6.4 bersama dengan *Bayesian Network* (BN) dan *likelihood ratio test* dapat digunakan untuk mengidentifikasi dan menguji kesamaan pola dokumen teks dengan memperhatikan pola struktur *term* (munculnya tiga *term* pertama) di setiap kalimat dalam dokumen teks yang dibandingkan. Hasil

dengan perhitungan *Bayesian Network* (BN) dan *likelihood ratio test* diperoleh 100% sesuai skenario dokumen asli.

Dari ketiga metode identifikasi dan pengujian kesamaan pola dokumen yang dihasilkan, cara identifikasi dan pengujian pola berdasarkan struktur *term* (munculnya tiga *term* pertama) di setiap kalimat dalam dokumen teks lebih akurat, dikarenakan penggunaan tiga *term* pertama dengan representasi penghitungan probabilitas bersamanya lebih bisa mewakili pola kalimat.

8.2 Saran

Berdasarkan hasil penelitian ini, maka beberapa saran yang direkomendasikan adalah sebagai berikut:

1. Hasil penelitian ini dapat digunakan sebagai langkah awal pendeteksian kesamaan dengan pola munculnya *term* dari kalimat dalam dokumen teks, sehingga pada masa yang akan datang dapat dilakukan pengembangan model untuk mendapatkan model pendeteksian pola dokumen teks yang terbaik.
2. Hasil penelitian ini dapat dilanjutkan dengan pembuatan *software* sebagai alat pendeteksian kesamaan dokumen teks dengan pola munculnya *term* dari kalimat dalam dokumen teks.

DAFTAR PUSTAKA

Ardiansyah, A. (2011), *Pengembangan Aplikasi Pendeteksi Plagiarisme Menggunakan Metode Latent Semantic Analysis (LSA), (Studi Kasus Plagiarisme Karya Ilmiah Berbahasa Indonesia)*, Universitas Pendidikan Indonesia, Bandung.

Ardytha, L., Junta, Z. dan Abu, S. (2013), *Algoritma Latent Semantic Analysis (LSA) Pada Peringkas Dokumen Otomatis Untuk Proses Clustering Dokumen*, SEMANTIK, Semarang.

Beeferman, D., Berger, A. dan Lafferty, J. (1999), "Statistical models for text segmentation", *Machine Learning, Special Issue on Natural Language Learning*, Vol. 34, hal. 177-210.

Box, G.E.P dan Tiao, G.C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Ben-Gal, I., Ruggeri, F., Faltin F. dan Kenett, R. (2007), *Bayesian Networks*, Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons, Inc., New York.

Bela, G. dan Meuschke, N. (2011), "Citation Pattern Matching Algorithms for Citationbased Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence", *Proceeding of the 11th ACM Symposium on Document Engineering*, Mountain View, CA, New York, pp. 249-258.

Congdon, P.D. (2006), *Bayesian Statistical Modelling*, 2nd edition, John iley & Sons, England.

Chakravart, I.M., Laha, R.G. dan Roy, J. (1967), *Handbook of methods of applied statistics*, John Wiley and Sons, New York.

Carlin, B.P. dan Chib, S. (1995), "Bayesian model choice via Markov Chain Monte Carlo methods", *Journal of the Royal Statistical Society*, Ser. B, 57(3): 473-484.

Chow, C.K. dan Liu, C.N. (1968), "Approximating discrete probability distributions with dependence trees", *IEEE Transactions on Information Theory*, Vol. IT-14, No. 3, hal. 462-467.

Cooper, G.F. (1990), "The computational complexity of probabilistic inference using Bayesian belief networks", *Artificial Intelligence*, Vol. 42, No. 2-3, hal. 393-405.

Cano, R., Sordo, C. dan Jose, MG. (2004), "Applications of Bayesian Networks in Meteorology", *Advances in Bayesian Networks*, Gámez et al. eds., hal. 309-327.

Cofino, A.S., Cano, R., Sordo, C. dan Gutierrez, J.M. (2002), "Bayesian Networks for Probabilistic Weather Prediction", *Proceedings of the 15th European Conference on Artificial Intelligence*, IOS Press, hal. 695-700.

Cooper, G.F. dan Herskovits F. (1992), "A Bayesian Method for The Induction of Probabilistic Networks from Data", *Machine Learning Journal* , Vol. 9, hal. 308-347.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G. dan Harshman, R. (1990), "Indexing by Latent Semantic Analysis", *Journal of the American Society of Information Science*, Vo. 41, No. 6, hal. 391-407.

Dumais, S.T. (2004), "Latent Semantic Analysis", *Annual Review of Information Science and Technology*, Vol. 30, hal. 188-230.

Gelman, A., Carlin, J.B., Stern, H.S. dan Rubin, D.B. (2004), *Bayesian Data Analysis*, 2nd edition, Chapman & Hall, Florida.

Ghanem, M., Guo, Y., Lodhi, H. dan Zhang, Y. (2002), "Automatic scientific text classification using local patterns", *ACM SIGKDD Explore News*, Vol. 4, No. 2, hal. 95-96.

Grim, J., Novoviˇcov´a dan Somol, P. (2008), "Structural Poisson mixtures for classification of documents", *Proceedings of the 19th International Conference on Pattern Recognition (ICPR '08)*, hal. 1-4.

Hansen, P.C. (1987), "The truncated SVD as a method for regularization", *BIT Numerical Mathematics*, Vol. 27, No. 4, hal. 534-553.

Helm, L. (1996), *The future of software may lie in the obscure theories of an 18th century cleric named Thomas Bayes*, Times Staff Writer, Los Angeles Times.

Joachims, T. (1997), "A probabilistic analysis of the Rocchio algorithm with TF-IDF for text categorization", *Proceedings of the 40th International Conference on Machine Learning*, hal. 143-151.

Jensen, F.V. (1996), *An Introduction to Bayesian Networks*, Springer-Verlag, NewYork

Kusner, M.J., Sun, Y., Kolkin, N.I. dan Weinberger, K.Q. (2015), "From Word Embeddings To Document Distances", *Proceedings of the 32 nd International Conference on Machine Learning*, Vol. 37, hal. 957-966.

King, R., Morgan, B.J.T, Gimenez, O., dan Brooks, S.P. (2010), *Bayesian Analysis for Population Ecology*, Chapman & Hall/CRC, USA.

Kasim, S. (2012), *Pembuatan Aplikasi Untuk Mendeteksi Plagiarisme dengan Metode Latent Semantic Analysis*, Tugas Akhir, Jurusan Teknik Informatika, Universitas Surabaya.

Kevin, B.K. dan Ann, E.N. (2011), *Bayesian Artificial Intelligence Second Edition*, Computer Science and Data Analysis Series, CRC Press, New York.

Kass R.E. dan Raftery, A.E. (1995), "Bayes Factors", *Journal of the American Statistical Association*, Vol. 90, No. 430, hal. 773-795.

Kintsch, W. (2001). "Predication", *Cognitive Science*, Vol. 25, hal. 173–202.

Lehman, E.L. dan Romano, J.P. (2005), *Testing Statistical Hypotheses 3rd Edition*, Springer, New York.

Lopes, R.H.C., Reid, I., dan Hobson, P.R. (2007), "The two-dimensional Kolmogorov-Smirnov test", *XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research*, Amsterdam.

Law, A.M., dan Kelton, W.D. (2000), *Simulation Modeling and Analysis*, McGraw-Hill International Series, Singapore.

Landauer, T.K., dan Dumais, S.T. (1997), "A solution to Plato's problem : The Latent Semantic Analysis theory of the acquisition, induction and

representaion of knowledge", *Psychological Review*, Vol. 104, No. 2, hal. 211-240.

Liyanto, (2011): https://liyantanto.wordpress.com/2011/06/28/pencarian-dengan-metode-vektor-space-model-vsm/

Landauer, T.K., Foltz, P.W. dan Laham, D. (1998), "Introduction to Latent Semantic Analysis", *Discourse Processes*, Vol. 25, No. 2-3, hal. 259-284.

Munir, R., (2015), Aplikasi Aljabar Vektor pada Sistem Temu-balik Informasi, Bahan Kuliah IF2123 Aljabar Geometri, Program Studi Teknik Informatika Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung.

Mittal, A., Ashraf K. dan Tele, T. (2007), *Bayesian Network Technologies: Applications and Graphical Models*, IGI Publishing, New York.

Ntzoufras, I. (2009), *Bayesian Modeling Using WinBUGS*. Wiley, New Jersey, USA.

Ozgur, L. dan Gungor, T. (2010), "Text classification with the support of pruned dependency patterns", *Journal Pattern Recognition Letters*, Vol. 31, No. 12, hal. 1598-1607.

Ogura, H., Amano, H. dan Kondo, M. (2013), *Gamma-Poisson Distribution Model for Text Categorization*, Faculty of Arts and Sciences, Showa University, 4562, Kamiyashida, Japan.

Peacock, J.A. (1983), "Two-dimensional goodness-of-fit testing in astronomy", *Journal Monthly Notices of the Royal Astronomical Society*, Vol. 202, hal. 615-627.

Pollino, CA. (2006), *Quantifying Ecological Risks using Bayesian Networks: Modelling in an Uncertain World*, MRC Ecological Risk Assessment Training Course.

Pawan G., Laxmidhar B. dan McGinnity, TM. (2008), "Application of Bayesian Framework in Natural Language Understanding", *IETE TECHNICAL REVIEW*, Vol. 25, hal. 251 -269.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan and Kaufmann, San Mateo, CA.

Peter, R., Shivapratap, G., Divya, G. dan Soma, KP. (2009), "Evaluation of SVD and NMF Methods for Latent Semantic Analysis", *International Journal of Recent Trends in Engineering*, Vol. 1, hal. 308-310.

Ribeiro-Neto, B.A. dan Muntz, R.R. (1996), "A belief network model for IR", *Proceedings of the 19th ACM SIGIR Conference*, hal. 253-260.

Salmuasih (2013), *Perancangan system deteksi plagiat pada dokumen teks dengan konsep similarity menggunakan Algoritma Rabin Karp*, Tugas Akhir Sekolah Tinggi Manajemen Informatika dan Komputer, AMIKOM, Yogyakarta.

Stein, B. dan Eissen, SM. (2006), "Near Similarity Search and Plagiarism Analysis", *Papers from the 29th Annual Conference of the German Classification Society (GfKl) Magdeburg*, hal. 430-437.

Soehardjoepri, Iriawan N., Ulama, B.S.S. dan Irhamah (2013), "On the Text Documents Pattern Recognition Using Latent Semantic Analysis and Kolmogorov-Smirnov Test", *Proceedings South East Asian Conference on Mathematics and Its Applications*, Department of Mathematics, FMIPA-ITS, Surabaya, hal. AM-24.

Soehardjoepri, Iriawan, N., Ulama, B.S.S. dan Irhamah (2015), "Identifying Text Document Pattern For Two Terms Appearances VIA Latent Semantic Analysis (LSA) Method And Term Distance Between Two Documents", *Journal of Theoretical and Applied Information Technology*, Vol. 79, No. 2, hal. 322-329.

Soehardjoepri, Iriawan, N., Ulama, B.S.S. dan Irhamah (2016), "On The Identification Of The Structural Pattern Of Terms Occurrence In a Document Using Bayesian Network", *Journal of Theoretical and Applied Information Technology*, Vol. 92, No. 2, hal. 253-264.

Sastroasmoro, S. (2007), "Beberapa Catatan Tentang *Plagiarisme*", *Majalah Kedokteran Indonesia*, Vol. 57, No. 8, hal. 239-244.

Sahoo, P. (2013), *Probability and Mathematical Statistics*, Departement of Mathematics, University of Louisville, KY 40292 USA.

Somayasa (2008), *Diktat Kuliah Statistika Matematika I*, Kendari: Universitas Halu Oleo.

Sentosa, J. (2015), *Aplikasi Model Ruang Vektor dan Matriks untuk Mndeteksi Adanya Plagiarisme*, makalah IF2123 Aljabar Geometri, Informatika ITB

Triawati, C. (2009), *Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia*, Skripsi, Institut Teknologi Telkom. Bandung.

Turtle, H.R. dan Croft, W.B. (1991), "Efficient probabilistic inference for text retrieval", *Proceedings of the RIA0'91 Conference*, hal. 644-661.

Thalib, F. dan Kusumawati, R. (2010), *Pembuatan Program Aplikasi Untuk Pendeteksian Kemiripan Dokumen Teks Dengan Algoritma Smith-Waterman*, Universitas Gunadarma, Depok-Indonesia.

Tan, P., Steinbach, M. dan Kumar, V. (2006), *Introduction to Data Mining*, Pearson Addison-Wesley

Winoto, H. (2012), *Deteksi Kemiripan Isi Dokumen Teks Menggunakan Algoritma Levenshtein Distance*, Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim, Malang.

Wong, S.K.M. dan Butz, C.J. (2000), "A Bayesian approach to user profiling in Information Retrieval", *In Technology Letters*, Vol. 4, No. 1, hal. 50-56.

Yoga, K.V.Y. (2012), "Pengembangan Aplikasi Pendeteksian Plagiarisme Pada Dokumen Teks Menggunakan Algoritma Rabin-Karp", *Kumpulan Artikel Mahasiswa Pendidikan Teknik Informatika (KARMAPATI)*, Vol. 1, No. 4, hal. 429-443.

Lampiran 1. Dokumen Asli, Hasil *Filtering* dan Hasil *Stemming*

Dokumen 1 (Dok-1)

Akustik adalah ilmu yang mempelajari perilaku bunyi dan sangat penting pada ruangan. Dinding yang keras dan polos dari sebuah ruangan akan memantulkan bunyi dan membuat ruangan tersebut bergema. Ruangan yang kecil akan terbantu mencegah hal ini bila ada bahan pada dinding dan langit-langit yang menyerap bunyi. Tirai dan karpet yang tebal juga akan membantu. Pada ruangan yang besar seperti gedung konser, diperlukan permukaan yang halus dan keras di belakang para peminat atau penyanyi untuk membantu membawa bunyi ke arah penonton, dan bahan yang menyerap bunyi di belakang gedung untuk mencegah gema.
(**91 kata**)

Hasil Fitering (Buang kata yang tidak penting)
akustik ilmu mempelajari perilaku bunyi ruangan dinding keras polos ruangan memantulkan bunyi membuat ruangan bergema ruangan kecil mencegah bahan dinding langit-langit menyerap bunyi tirai karpet tebal ruangan besar gedung konser diperlukan permukaan halus keras peminat penyanyi membawa bunyi arah penonton bahan menyerap bunyi gedung mencegah gema
(**46 kata**)

Hasil Stemming (Buang imbuhan)
akustik ilmu pelajari perilaku bunyi ruang dinding keras polos ruang pantul bunyi buat ruang gema ruang kecil cegah bahan dinding langit-langit serap bunyi tirai karpet tebal ruang besar gedung konser perlu muka halus keras minat penyanyi bawa bunyi arah penonton bahan serap bunyi gedung cegah gema
(**46 kata**)

Dokumen 2 (Dok-2)

Ruangan yang kecil akan terbantu mencegah hal ini bila ada bahan pada dinding dan langit-langit yang menyerap bunyi. Pada ruangan yang besar seperti gedung konser, diperlukan permukaan yang halus dan keras di belakang para peminat atau penyanyi untuk membantu membawa bunyi ke arah penonton, dan bahan yang menyerap bunyi di belakang gedung untuk mencegah gema. Akustik adalah ilmu yang mempelajari perilaku bunyi dan sangat penting pada ruangan. Dinding yang keras dan polos dari sebuah ruangan akan memantulkan bunyi dan membuat ruangan tersebut bergema. Tirai dan karpet yang tebal juga akan membantu.
(**91 kata**)

Lampiran 1. (lanjutan)

Hasil Fitering (Buang kata yang tidak penting)
ruangan kecil mencegah bahan dinding langit-langit menyerap bunyi ruangan besar gedung konser diperlukan permukaan halus keras peminat penyanyi membawa bunyi arah penonton bahan menyerap bunyi gedung mencegah gema akustik ilmu mempelajari perilaku bunyi ruangan dinding keras polos ruangan memantulkan bunyi membuat ruangan bergema tirai karpet tebal
(46 kata)

Hasil Stemming (Buang imbuhan)
ruang kecil cegah bahan dinding langit-langit serap bunyi ruang besar gedung konser perlu muka halus keras peminat penyanyi bawa bunyi arah penonton bahan serap bunyi gedung cegah gema akustik ilmu pelajari perilaku bunyi ruang dinding keras polos ruang pantul bunyi buat ruang gema tirai karpet tebal
(46 kata)

Dokumen 3 (Dok-3)

Sebagaimana sebuah daerah pada umumnya, Lampung memiliki beraneka ragam jenis musik, mulai dari jenis tradisional hingga modern, yang mengadopsi kebudayaan musik global. Adapun jenis musik yang masih bertahan hingga sekarang adalah Klasik Lampung. Jenis musik ini biasanya diiringi oleh alat musik gambus dan gitar akustik. Jenis musik ini merupakan perpaduan budaya Islam dan budaya asli itu sendiri. Beberapa kegiatan festival diadakan untuk mengembangkan budaya musik tradisional tanpa harus khawatir akan kehilangan jati diri. Festival Krakatau contohnya, adalah sebuah Festival yang diadakan oleh Pemda Lampung yang bertujuan untuk mengenalkan Lampung kepada dunia luar dan sekaligus menjadi ajang promosi pariwisata.
(98 kata)

Hasil Fitering (Buang kata yang tidak penting)
daerah umumnya lampung memiliki beraneka ragam musik mulai tradisional hingga modern mengadopsi kebudayaan musik global musik masih bertahan hingga sekarang klasik lampung musik biasanya diiringi alat musik gambus gitar akustik musik budaya islam budaya asli sendiri kegiatan festival diadakan budaya musik tradisional tanpa harus khawatir kehilangan jati diri festival krakatau festival diadakan pemda lampung bertujuan mengenalkan lampung dunia luar menjadi ajang promosi pariwisata.
(63 kata)

Lampiran 1. (lanjutan)

Hasil Stemming (Buang imbuhan)
daerah umum lampung milik aneka ragam musik mulai tradisional hingga modern adopsi budaya musik global musik masih tahan hingga sekarang klasik lampung musik biasa diiringi alat musik gambus gitar akustik musik budaya islam budaya asli sendiri kegiatan festival ada budaya musik tradisional tanpa harus khawatir hilang jati diri festival krakatau festival ada pemda lampung tujuan kenal lampung dunia luar menjadi ajang promosi pariwisata
(63 kata)

Dokumen 4 (Dok-4)

Beragam jenis musik dimiliki oleh daerah Lampung, sebagaimana sebuah daerah pada umumnya, mulai dari jenis tradisional hingga modern, yang mengadopsi kebudayaan musik global. Klasik Lampung adalah jenis musik yang masih bertahan hingga sekarang. Alat musik Gambus dan gitar akustik biasanya mengiringi jenis musik ini. Jenis musik ini merupakan perpaduan budaya Islam dan budaya asli itu sendiri. Budaya musik tradisional dikembangkan tanpa harus khawatir akan kehilangan jati diri dengan cara mengadakan beberapa kegiatan festival. Pemda Lampung mengadakan sebuah festival yang bertujuan untuk mengenalkan Lampung kepada dunia luar dan sekaligus menjadi ajang promosi pariwisata, dinamakan Festival Krakatau.
(95 kata)

Hasil Fitering (Buang kata yang tidak penting)
beragam musik dimiliki daerah lampung daerah umumnya mulai tradisional hingga modern mengadopsi kebudayaan musik global klasik lampung musik masih bertahan hingga sekarang alat musik gambus gitar akustik biasanya mengiringi musik musik perpaduan budaya islam budaya asli sendiri budaya musik tradisional dikembangkan harus khawatir kehilangan jati diri cara mengadakan kegiatan festival pemda lampung mengadakan festival bertujuan mengenalkan lampung dunia luar menjadi ajang promosi pariwisata festival krakatau
(65 Kata)

Hasil Stemming (Buang imbuhan)
ragam musik milik daerah lampung daerah umum mulai tradisional hingga modern adopsi budaya musik global klasik lampung musik masih tahan hingga sekarang alat musik gambus gitar akustik biasa iring musik musik padu budaya islam budaya asli sendiri budaya musik tradisional kembang harus khawatir hilang jati diri cara ada kegiatan festival pemda lampung ada festival tujuan kenal lampung dunia luar jadi ajang promosi pariwisata festival krakatau
(65 kata)

Lampiran 1. (lanjutan)

Dokumen 5 (Dok-5)

Beragam jenis musik dimiliki oleh daerah Lampung, sebagaimana sebuah daerah pada umumnya, mulai dari jenis tradisional hingga modern, yang mengadopsi kebudayaan musik global. Klasik Lampung adalah jenis musik yang masih bertahan hingga sekarang. Alat musik Gambus dan gitar akustik biasanya mengiringi jenis musik ini. Jenis musik ini merupakan perpaduan budaya Islam dan budaya asli itu sendiri. Budaya musik tradisional dikembangkan tanpa harus khawatir akan kehilangan jati diri dengan cara mengadakan beberapa kegiatan festival. Pemda Lampung mengadakan sebuah festival yang bertujuan untuk mengenalkan Lampung kepada dunia luar dan sekaligus menjadi ajang promosi pariwisata, dinamakan Festival Krakatau. Akustik adalah ilmu yang mempelajari perilaku bunyi dan sangat penting pada ruangan. Dinding yang keras dan polos dari sebuah ruangan akan memantulkan bunyi dan membuat ruangan tersebut bergema. Ruangan yang kecil akan terbantu mencegah hal ini bila ada bahan pada dinding dan langit-langit yang menyerap bunyi. Tirai dan karpet yang tebal juga akan membantu. Pada ruangan yang besar seperti gedung konser, diperlukan permukaan yang halus dan keras di belakang para peminat atau penyanyi untuk membantu membawa bunyi ke arah penonton, dan bahan yang menyerap bunyi di belakang gedung untuk mencegah gema. (186 kata)

Hasil Fitering (Buang kata yang tidak penting)

beragam musik dimiliki daerah lampung daerah umumnya mulai tradisional hingga modern mengadopsi kebudayaan musik global klasik lampung musik bertahan hingga sekarang alat musik gambus gitar akustik biasanya mengiringi musik musik perpaduan budaya islam budaya asli sendiri budaya musik tradisional dikembangkan harus khawatir kehilangan jati diri cara mengadakan kegiatan festival pemda lampung mengadakan festival bertujuan mengenalkan lampung dunia luar ajang promosi pariwisata festival krakatau akustik ilmu mempelajari perilaku bunyi penting ruangan dinding keras polos ruangan memantulkan bunyi membuat ruangan bergema ruangan kecil terbantu mencegah ada bahan dinding langit-langit menyerap bunyi tirai karpet tebal juga ruangan besar gedung konser diperlukan permukaan halus keras peminat penyanyi membawa bunyi arah penonton bahan menyerap bunyi gedung mencegah gema.
(113 kata)

Hasil Stemming (Buang imbuhan)

ragam musik milik daerah lampung daerah umum mulai tradisional hingga modern adopsi budaya musik global klasik lampung musik tahan hingga sekarang alat musik gambus gitar akustik biasa iring musik musik padu budaya islam budaya asli sendiri budaya musik tradisional kembang harus khawatir hilang jati diri cara ada kegiatan festival pemda lampung ada festival tujuan kenal lampung dunia luar ajang promosi pariwisata festival krakatau akustik ilmu pelajari perilaku bunyi penting ruang dinding keras polos ruang pantul bunyi buat ruang gema ruang kecil bantu cegah ada bahan dinding langit-langit serap bunyi tirai karpet tebal juga ruang besar gedung konser perlu muka halus keras minat penyanyi bawa bunyi arah penonton bahan serap bunyi gedung cegah gema.
(113 kata)

Lampiran 1. (lanjutan)

Dokumen 6 (Dok-6)

Kita memperoleh apa yang kita inginkan melalui negosiasi. Mulai dari bangun pagi, mungkin kita harus mengambil kesepakatan siapa yang harus menggunakan kamar mandi terlebih dahulu, kemudian apakah sopir harus mengantar isteri anda atau anda terlebih dahulu. Demikian pula di kantor misalnya kita melakukan negosiasi dalam rapat direksi, rapat staf, bahkan untuk menentukan di mana akan makan siang kita harus bernegosiasi dengan rekan sekerja kita. Jadi kita semua pada dasarnya adalah negosiator. Beberapa dari kita melakukannya dengan baik, sedangkan sebagian lagi tidak pernah memenangkan negosiasi. Sebagian kita hanya menjadi pengikut atau selalu mengikuti dan mengakomodasi kepentingan orang lain. Negosiasi dilakukan oleh semua manusia yang berinteraksi dengan manusia lainnya. Mulai dari anak kecil sampai orang tua, semua lapisan dari kalangan sosial terbawah sampai dengan kaum elit di kalangan atas. Negosiasi dilakukan mulai dari rumah, sekolah, kantor, dan semua aspek kehidupan kita. Oleh karena itu penting bagi kita dalam rangka mengembangkan dan mengelola diri (manajemen diri), untuk dapat memahami dasar-dasar, prinsip dan teknik-teknik bernegosiasi sehingga kita dapat melakukan negosiasi serta membangun relasi yang jauh lebih efektif dan lebih baik dengan siapa saja. (179 kata)

Hasil Fitering (Buang kata yang tidak penting)

memperoleh apa inginkan melalui negosiasi mulai bangun pagi mungkin harus mengambil kesepakatan harus menggunakan kamar mandi terlebih dahulu sopir harus mengantar isteri terlebih dahulu pula kantor melakukan negosiasi rapat direksi rapat staf menentukan makan siang harus bernegosiasi rekan sekerja jadi semua dasarnya negosiator melakukannya baik lagi pernah memenangkan negosiasi hanya menjadi pengikut selalu mengikuti mengakomodasi kepentingan orang negosiasi dilakukan semua manusia berinteraksi manusia lainnya mulai anak kecil orang tua semua lapisan kalangan sosial terbawah kaum elit kalangan atas negosiasi dilakukan mulai rumah sekolah kantor semua aspek kehidupan penting dalam rangka mengembangkan mengelola diri manajemen diri dapat memahami dasar-dasar prinsip teknik-teknik bernegosiasi dapat melakukan negosiasi membangun relasi jauh lebih efektif lebih baik siapa saja. (113 kata)

Hasil Stemming (Buang imbuhan)

peroleh apa ingin melalui negosiasi mulai bangun pagi mungkin harus ambil sepakat harus guna kamar mandi lebih dahulu sopir harus antar isteri lebih dahulu pula kantor lakukan negosiasi rapat direksi rapat staf tentukan makan siang harus negosiasi rekan sekerja jadi semua dasar negosiator lakukan baik lagi pernah menang negosiasi hanya jadi ikut selalu ikut akomodasi penting orang negosiasi lakukan semua manusia interaksi manusia lain mulai anak kecil orang tua semua lapisan kalangan sosial bawah kaum elit kalangan atas negosiasi lakukan mulai rumah sekolah kantor semua aspek hidup penting dalam rangka kembangkan kelola diri manajemen diri dapat paham dasar-dasar prinsip teknik-teknik negosiasi dapat lakukan negosiasi bangun relasi jauh lebih efektif lebih baik siapa saja. (113 kata)

Lampiran 2. Tampilan Koding *Term* Untuk Setiap Dokumen

Dokumen-1 :

T19 T2 T1 T9
Akustik adalah ilmu yang mempelajari perilaku bunyi dan sangat penting pada ruangan. Dinding
T7 T1 T2 T1
yang keras dan polos dari sebuah ruangan akan memantulkan bunyi dan membuat ruangan
T3 T1 T4 T18
tersebut bergema. Ruangan yang kecil akan terbantu mencegah hal ini bila ada bahan pada
T9 T8 T8 T6 T2
dinding dan langit-langit yang menyerap bunyi. Tirai dan karpet yang tebal juga akan membantu.
T1 T5 T7
Pada ruangan yang besar seperti gedung konser, diperlukan permukaan yang halus dan keras di
T2
belakang para peminat atau penyanyi untuk membantu membawa bunyi ke arah penonton, dan
T18 T6 T2 T5 T4 T3
bahan yang menyerap bunyi di belakang gedung untuk mencegah gema.

Dokumen-2 :

T1 T4 T18 T9 T8 T8
Ruangan yang kecil akan terbantu mencegah hal ini bila ada bahan pada dinding dan langit-langit
T6 T2 T1 T5
yang menyerap bunyi. Pada ruangan yang besar seperti gedung konser, diperlukan permukaan
T7 T2
yang halus dan keras di belakang para peminat atau penyanyi untuk membantu membawa bunyi
T18 T6 T2 T5 T4 T3
ke arah penonton, dan bahan yang menyerap bunyi di belakang gedung untuk mencegah gema.
T19 T2 T1 T9
Akustik adalah ilmu yang mempelajari perilaku bunyi dan sangat penting pada ruangan. Dinding
T7 T1 T2 T1
yang keras dan polos dari sebuah ruangan akan memantulkan bunyi dan membuat ruangan
T3
tersebut bergema. Tirai dan karpet yang tebal juga akan membantu.

Lampiran 2. (lanjutan)

Dokumen-3 :

T17 T12 T10
Sebagaimana sebuah daerah pada umumnya, Lampung memiliki beraneka ragam jenis musik,
T16 T15 T11 T10
mulai dari jenis tradisional hingga modern, yang mengadopsi kebudayaan musik global. Adapun
T10 T15 T12 T10
jenis musik yang masih bertahan hingga sekarang adalah Klasik Lampung. Jenis musik ini
T10 T19 T10
biasanya diiringi oleh alat musik gambus dan gitar akustik. Jenis musik ini merupakan perpaduan
T11 T11 T13 T14
budaya Islam dan budaya asli itu sendiri. Beberapa kegiatan festival diadakan untuk
T11 T10 T16
mengembangkan budaya musik tradisional tanpa harus khawatir akan kehilangan jati diri.
T13 T13 T14 T12
Festival Krakatau contohnya, adalah sebuah Festival yang diadakan oleh Pemda Lampung yang
T12
bertujuan untuk mengenalkan Lampung kepada dunia luar dan sekaligus menjadi ajang promosi

pariwisata.

Dokumen-4 :

T10 T17 T12 T17
Beragam jenis musik dimiliki oleh daerah Lampung, sebagaimana sebuah daerah pada
T16 T15 T11 T10
umumnya, mulai dari jenis tradisional hingga modern, yang mengadopsi kebudayaan musik
T12 T10 T15 T10
global. Klasik Lampung adalah jenis musik yang masih bertahan hingga sekarang. Alat musik
T19 T10 T10
Gambus dan gitar akustik biasanya mengiringi jenis musik ini. Jenis musik ini merupakan
T11 T11 T11 T10 T16
perpaduan budaya Islam dan budaya asli itu sendiri. Budaya musik tradisional dikembangkan
T14
tanpa harus khawatir akan kehilangan jati diri dengan cara mengadakan beberapa kegiatan
T13 T12 T14 T13
festival. Pemda Lampung mengadakan sebuah festival yang bertujuan untuk mengenalkan
T12 T13
Lampung kepada dunia luar dan sekaligus menjadi ajang promosi pariwisata, dinamakan Festival

Krakatau.

Lampiran 2. (lanjutan)

Dokumen-5 :

T10 T17 T12 T17
Beragam jenis musik dimiliki oleh daerah Lampung, sebagaimana sebuah daerah pada
T16 T15 T11 T10
umumnya, mulai dari jenis tradisional hingga modern, yang mengadopsi kebudayaan musik
T12 T10 T15 T10
global. Klasik Lampung adalah jenis musik yang masih bertahan hingga sekarang. Alat musik
T19 T10 T10
Gambus dan gitar akustik biasanya mengiringi jenis musik ini. Jenis musik ini merupakan
T11 T11 T11 T10 T16
perpaduan budaya Islam dan budaya asli itu sendiri. Budaya musik tradisional dikembangkan
T14
tanpa harus khawatir akan kehilangan jati diri dengan cara mengadakan beberapa kegiatan
T13 T12 T14 T13
festival. Pemda Lampung mengadakan sebuah festival yang bertujuan untuk mengenalkan
T12 T13
Lampung kepada dunia luar dan sekaligus menjadi ajang promosi pariwisata, dinamakan Festival
T19 T2
Krakatau. Akustik adalah ilmu yang mempelajari perilaku bunyi dan sangat penting pada
T1 T9 T7 T1 T2
ruangan. Dinding yang keras dan polos dari sebuah ruangan akan memantulkan bunyi dan
T1 T3 T1 T4
membuat ruangan tersebut bergema. Ruangan yang kecil akan terbantu mencegah hal ini bila ada
T18 T9 T8 T8 T6 T2
bahan pada dinding dan langit-langit yang menyerap bunyi. Tirai dan karpet yang tebal juga akan
T1 T5
membantu. Pada ruangan yang besar seperti gedung konser, diperlukan permukaan yang halus
T7 T2
dan keras di belakang para peminat atau penyanyi untuk membantu membawa bunyi ke arah
T18 T6 T2 T5 T4 T3
penonton, dan bahan yang menyerap bunyi di belakang gedung untuk mencegah gema.

Lampiran 2. (lanjutan)

Dokumen-6 :

Kita memperoleh apa yang kita inginkan melalui negosiasi. Mulai dari bangun pagi, mungkin kita harus mengambil kesepakatan siapa yang harus menggunakan kamar mandi terlebih dahulu, kemudian apakah sopir harus mengantar isteri anda atau anda terlebih dahulu. Demikian pula di kantor misalnya kita melakukan negosiasi dalam rapat direksi, rapat staf, bahkan untuk menentukan di mana akan makan siang kita harus bernegosiasi dengan rekan sekerja kita. Jadi kita semua pada dasarnya adalah negosiator. Beberapa dari kita melakukannya dengan baik, sedangkan sebagian lagi tidak pernah memenangkan negosiasi. Sebagian kita hanya menjadi pengikut atau selalu mengikuti dan mengakomodasi kepentingan orang lain. Negosiasi dilakukan oleh semua manusia yang berinteraksi dengan manusia lainnya. Mulai dari anak kecil sampai orang tua, semua lapisan dari kalangan sosial terbawah sampai dengan kaum elit di kalangan atas. Negosiasi dilakukan mulai dari rumah, sekolah, kantor, dan semua aspek kehidupan kita.

Oleh karena itu penting bagi kita dalam rangka mengembangkan dan mengelola diri (manajemen diri), untuk dapat memahami dasar-dasar, prinsip dan teknik-teknik bernegosiasi sehingga kita dapat melakukan negosiasi serta membangun relasi yang jauh lebih efektif dan lebih baik dengan

siapa saja.

Lampiran 3. Perhitungan jarak (*d*) setiap pasangan *term* dari setiap dua dokumen

Dok-1					Dok-2					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{19}	1	T_2	1	1	T_1	3	T_4	3	1	2.83
T_9	2	T_7	2	2	T_1	3	T_5	4	2	2.24
T_1	3	T_4	3	3	T_{19}	1	T_2	1	3	2.83
T_1	3	T_5	4	4	T_9	2	T_7	2	4	2.24

Dok-1					Dok-3					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{19}	1	T_2	1	1	T_{17}	4	T_{12}	5	1	5.00
T_9	2	T_7	2	2	T_{10}	5	T_{15}	6	2	5.00
T_1	3	T_4	3	3	T_{10}	5	T_{10}	7	3	4.47
T_1	3	T_5	4	4	T_{10}	5	T_{11}	8	4	4.47
T_1	3	T_5	4	5	T_{13}	6	T_{14}	9	5	5.83
T_1	3	T_5	4	6	T_{13}	6	T_{13}	10	6	6.71

Dok-1					Dok-4					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{19}	1	T_2	1	1	T_{10}	4	T_{17}	5	1	5.00
T_9	2	T_7	2	2	T_{12}	5	T_{10}	6	2	5.00
T_1	3	T_4	3	3	T_{10}	4	T_{19}	7	3	4.12
T_1	3	T_5	4	4	T_{10}	4	T_{11}	8	4	4.12
T_1	3	T_5	4	5	T_{11}	6	T_{10}	6	5	3.61
T_1	3	T_5	4	6	T_{12}	5	T_{14}	9	6	5.39

Dok-1					Dok-5					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{19}	1	T_2	1	1	T_{10}	4	T_{17}	5	1	5.00
T_9	2	T_7	2	2	T_{12}	5	T_{10}	6	2	5.00
T_1	3	T_4	3	3	T_{10}	4	T_{19}	7	3	4.12
T_1	3	T_5	4	4	T_{10}	4	T_{11}	8	4	4.12
T_1	3	T_5	4	5	T_{11}	6	T_{10}	6	5	3.61
T_1	3	T_5	4	6	T_{12}	5	T_{14}	9	6	5.39
T_1	3	T_5	4	7	T_{19}	1	T_2	1	7	3.61
T_1	3	T_5	4	8	T_9	2	T_7	2	8	2.24
T_1	3	T_5	4	9	T_1	3	T_4	3	9	1.00
T_1	3	T_5	4	10	T_1	3	T_5	4	10	0.00

Lampiran 3. (lanjutan)

Dok-1					Dok-6					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{19}	1	T_2	1	1	T_{30}	4	T_{21}	5	1	5.00
T_9	2	T_7	2	2	T_{24}	5	T_{20}	6	2	5.00
T_1	3	T_4	3	3	T_{20}	6	T_{23}	7	3	5.00
T_1	3	T_5	4	4	T_{20}	6	T_{27}	8	4	5.00
T_1	3	T_5	4	5	T_{26}	7	T_{29}	9	5	6.40
T_1	3	T_5	4	6	T_{22}	8	T_{22}	10	6	7.81

Dok-2					Dok-3					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_1	1	T_4	1	1	T_{17}	4	T_{12}	5	1	5.00
T_1	1	T_5	2	2	T_{10}	5	T_{15}	6	2	5.66
T_{19}	2	T_2	3	3	T_{10}	5	T_{10}	7	3	5.00
T_9	3	T_7	4	4	T_{10}	5	T_{11}	8	4	4.47
T_9	3	T_7	4	5	T_{13}	6	T_{14}	9	5	5.83
T_9	3	T_7	4	6	T_{13}	6	T_{13}	10	6	6.71

Dok-2					Dok-4					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_1	1	T_4	1	1	T_{10}	4	T_{17}	5	1	5.00
T_1	1	T_5	2	2	T_{12}	5	T_{10}	6	2	5.66
T_{19}	2	T_2	3	3	T_{10}	4	T_{19}	7	3	4.47
T_9	3	T_7	4	4	T_{10}	4	T_{11}	8	4	4.12
T_9	3	T_7	4	5	T_{11}	6	T_{10}	6	5	3.61
T_9	3	T_7	4	6	T_{12}	5	T_{14}	9	6	5.39

Lampiran 3. (lanjutan)

Dok-2					Dok-5					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_1	1	T_4	1	1	T_{10}	4	T_{17}	5	1	5.00
T_1	1	T_5	2	2	T_{12}	5	T_{10}	6	2	5.66
T_{19}	2	T_2	3	3	T_{10}	4	T_{19}	7	3	4.47
T_9	3	T_7	4	4	T_{10}	4	T_{11}	8	4	4.12
T_9	3	T_7	4	5	T_{11}	6	T_{10}	6	5	3.61
T_9	3	T_7	4	6	T_{12}	5	T_{14}	9	6	5.39
T_9	3	T_7	4	7	T_{19}	2	T_2	3	7	1.41
T_9	3	T_7	4	8	T_9	3	T_7	4	8	0.00
T_9	3	T_7	4	9	T_1	1	T_4	1	9	3.61
T_9	3	T_7	4	10	T_1	1	T_5	2	10	2.83

Dok-2					Dok-6					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_1	1	T_4	1	1	T_{30}	4	T_{21}	5	1	5.00
T_1	1	T_5	2	2	T_{24}	5	T_{20}	6	2	5.66
T_{19}	2	T_2	3	3	T_{20}	6	T_{23}	7	3	5.66
T_9	3	T_7	4	4	T_{20}	6	T_{27}	8	4	5.00
T_9	3	T_7	4	5	T_{26}	7	T_{29}	9	5	6.40
T_9	3	T_7	4	6	T_{22}	8	T_{22}	10	6	7.81

Dok-3					Dok-4					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{17}	1	T_{12}	1	1	T_{10}	2	T_{17}	7	1	6.08
T_{10}	2	T_{15}	2	2	T_{12}	4	T_{10}	3	2	2.24
T_{10}	2	T_{10}	3	3	T_{10}	2	T_{19}	8	3	5.00
T_{10}	2	T_{11}	4	4	T_{10}	2	T_{11}	4	4	0.00
T_{13}	3	T_{14}	5	5	T_{11}	5	T_{10}	3	5	2.83
T_{13}	3	T_{13}	6	6	T_{12}	4	T_{14}	5	6	1.41

Lampiran 3. (lanjutan)

Dok-3					Dok-5					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{17}	1	T_{12}	1	1	T_{10}	4	T_{17}	7	1	6.71
T_{10}	2	T_{15}	2	2	T_{12}	5	T_{10}	3	2	3.16
T_{10}	2	T_{10}	3	3	T_{10}	4	T_{19}	8	3	5.39
T_{10}	2	T_{11}	4	4	T_{10}	4	T_{11}	4	4	2.00
T_{13}	3	T_{14}	5	5	T_{11}	6	T_{10}	3	5	3.61
T_{13}	3	T_{13}	6	6	T_{12}	5	T_{14}	5	6	2.24
T_{13}	3	T_{13}	6	7	T_{19}	7	T_{2}	9	7	5.00
T_{13}	3	T_{13}	6	8	T_{9}	8	T_{7}	10	8	6.40
T_{13}	3	T_{13}	6	9	T_{1}	9	T_{4}	11	9	7.81
T_{13}	3	T_{13}	6	10	T_{1}	9	T_{5}	12	10	8.49

Dok-3					Dok-6					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{17}	1	T_{12}	1	1	T_{30}	4	T_{21}	7	1	6.71
T_{10}	2	T_{15}	2	2	T_{24}	5	T_{20}	8	2	6.71
T_{10}	2	T_{10}	3	3	T_{20}	6	T_{23}	9	3	7.21
T_{10}	2	T_{11}	4	4	T_{20}	6	T_{27}	10	4	7.21
T_{13}	3	T_{14}	5	5	T_{26}	7	T_{29}	11	5	7.21
T_{13}	3	T_{13}	6	6	T_{22}	8	T_{22}	12	6	7.81

Dok-4					Dok-5					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{10}	1	T_{17}	1	1	T_{10}	1	T_{17}	6	1	5.00
T_{12}	2	T_{10}	2	2	T_{12}	2	T_{10}	2	2	0.00
T_{10}	1	T_{19}	3	3	T_{10}	1	T_{19}	3	3	0.00
T_{10}	1	T_{11}	4	4	T_{10}	1	T_{11}	4	4	0.00
T_{11}	3	T_{10}	2	5	T_{11}	3	T_{10}	2	5	0.00
T_{12}	2	T_{14}	5	6	T_{12}	2	T_{14}	5	6	0.00
T_{12}	2	T_{14}	5	7	T_{19}	4	T_{2}	7	7	2.83
T_{12}	2	T_{14}	5	8	T_{9}	5	T_{7}	8	8	4.24
T_{12}	2	T_{14}	5	9	T_{1}	6	T_{4}	9	9	5.66
T_{12}	2	T_{14}	5	10	T_{1}	6	T_{5}	10	10	6.40

Lampiran 3. (lanjutan)

Dok-4					Dok-6					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{10}	1	T_{17}	1	1	T_{30}	4	T_{21}	6	1	5.83
T_{12}	2	T_{10}	2	2	T_{24}	5	T_{20}	7	2	5.83
T_{10}	1	T_{19}	3	3	T_{20}	6	T_{23}	8	3	7.07
T_{10}	1	T_{11}	4	4	T_{20}	6	T_{27}	9	4	7.07
T_{11}	3	T_{10}	2	5	T_{26}	7	T_{29}	10	5	8.94
T_{12}	2	T_{14}	5	6	T_{22}	8	T_{22}	11	6	8.49

Dok-5					Dok-6					*d*
Term	*X*	*Term*	*Y*	*Z*	*Term*	*X*	*Term*	*Y*	*Z*	
T_{10}	1	T_{17}	1	1	T_{30}	7	T_{21}	10	1	10.82
T_{12}	2	T_{10}	2	2	T_{24}	8	T_{20}	11	2	10.82
T_{10}	1	T_{19}	3	3	T_{20}	9	T_{23}	12	3	12.04
T_{10}	1	T_{11}	4	4	T_{20}	9	T_{27}	13	4	12.04
T_{11}	3	T_{10}	2	5	T_{26}	10	T_{29}	14	5	13.89
T_{12}	2	T_{14}	5	6	T_{22}	11	T_{22}	15	6	13.45
T_{19}	4	T_{2}	6	7	T_{22}	11	T_{22}	15	7	11.40
T_{9}	5	T_{7}	7	8	T_{22}	11	T_{22}	15	8	10.00
T_{1}	6	T_{4}	8	9	T_{22}	11	T_{22}	15	9	8.60
T_{1}	6	T_{5}	9	10	T_{22}	11	T_{22}	15	10	7.81

Lampiran 4. Perhitungan $\cos\theta$ untuk setiap pasangan Dokumen Uji

Term	D-1	D-2
ruang	5	5
bunyi	5	5
gema	2	2
cegah	2	2
gedung	2	2
serap	2	2
keras	2	2
langit	2	2
dinding	2	2
bahan	2	2
akustik	1	1

D-1.D-2 83.00

|D-1| 9.110434

|D-2| 9.110434

$\cos\theta =$ D-1.D-2/(|D-1||D2)

= 1

Term	D-1	D-3
ruang	0	7
bunyi	0	4
gema	0	4
cegah	0	3
gedung	0	2
serap	0	2
keras	0	2
langit	0	1
dinding	2	0
Akustik	1	1

D-1*D-3 1.00

|D-1| 2.236068 22.69361

|D-3| 10.14889

$\cos\theta =$ D-1.D-3/(|D-1||D3)

= 0,04

Term	D-1	D-4
ruang	0	7
bunyi	0	4
gema	0	4
cegah	0	3
gedung	0	2
serap	0	2
keras	0	2
langit	0	2
dinding	2	0
bahan	2	1
akustik	1	0

D-1*D-4 5.00

|D-1| 3

|D-4| 10.14889

$\cos\theta =$ D-1.D-4/(|D-1||D4|)

= 0,16

Lampiran 4. (Lanjutan)

Term	D-1	D-5
ruang	5	5
bunyi	5	5
gema	2	2
cegah	2	2
gedung	2	2
serap	2	2
keras	2	2
langit	2	2
dinding	2	2
musik	0	7
budaya	0	4
lampung	0	4
festival	0	3
ada	0	2
hingga	0	2
tradisional	0	2
daerah	0	2
bahan	2	2
akustik	1	2

D-1*D-5 83.00

|D-1| 9.110434

|D-5| 13.85641

$\cos\theta =$ D-1.D-5/(|D-1||D5|)

= 0,65

Term	D-1	D-6
negosiasi	0	8
lebih	0	4
dasar	0	3
ikut	0	2
kantor	0	2
rapat	0	2
orang	0	2
manusia	0	2
teknik	0	2
kalangan	0	2
bangun	0	2

D-1*D-6 0.00

|D-1| 0

|D-6| 11

$\cos\theta =$ D-1.D-6/(|D-1||D6|)

= Δ

Lampiran 4. (Lanjutan)

Term	D-2	D-3
ruang	0	7
bunyi	0	4
gema	0	4
cegah	0	3
gedung	0	2
serap	0	2
keras	0	2
langit	0	1
dinding	2	0
Akustik	1	1

D-2*D-3 1.00

|D-2| 2.236068

|D-3| 10.14889

$\cos\theta =$ D-2.D-3/(|D-2||D3|)

= 0,04

Term	D-2	D-4
ruang	0	7
bunyi	0	4
gema	0	4
cegah	0	3
gedung	0	2
serap	0	2
keras	0	2
langit	0	2
dinding	2	0
bahan	2	1
akustik	1	0

D-2*D-4 5.00

|D-2| 3

|D-4| 10.14889

$\cos\theta =$ D-2.D-4/(|D-2||D4|)

= 0,16

Term	D-2	D-5
ruang	5	5
bunyi	5	5
gema	2	2
cegah	2	2
gedung	2	2
serap	2	2
keras	2	2
langit	2	2
dinding	2	2
musik	0	7
budaya	0	4
lampung	0	4
festival	0	3
ada	0	2

D-2*D-5 84.00

|D-2| 9.110434

|D-5| 13.85641

$\cos\theta =$ D-2.D-5/(|D-2||D5|)

=0,66

Lampiran 4. (Lanjutan)

Term	D-2	D-5
hingga	0	2
tradisional	0	2
daerah	0	2
bahan	2	2
akustik	1	2

Term	D-2	D-6
negosiasi	0	8
lebih	0	4
dasar	0	3
ikut	0	2
kantor	0	2
rapat	0	2
orang	0	2
manusia	0	2
teknik	0	2
kalangan	0	2
bangun	0	2

D-2*D-6 0.00

|D-2| 0

|D-6| 11

$\cos\theta =$ D-2.D-6/(|D-2||D6|)
= Δ

Term	D-3	D-4
ruang	7	7
bunyi	4	4
gema	4	4
cegah	3	3
gedung	2	2
serap	2	2
keras	2	2
langit	1	2
dinding	0	0
Akustik	1	1

D-3*D-4 105.00

|D-3| 10.19804

|D-4| 10.34408

$\cos\theta =$ D-3.D-4/(|D-3||D4|)
= 0,99

Term	D-3	D-5
ruang	0	5
bunyi	0	5
gema	0	2
cegah	0	2
gedung	0	2
serap	0	2
keras	0	2

D-3*D-5 106.00

|D-3| 10.19804

|D-5| 13.85641

$\cos\theta =$ D-3.D-5/(|D-3||D5|)
= 0,75

Lampiran 4. (Lanjutan)

Term	D-3	D-5
langit	0	2
dinding	0	2
musik	7	7
budaya	4	4
lampung	4	4
festival	3	3
ada	2	2
hingga	2	2
tradisional	2	2
daerah	1	2
bahan	0	2
akustik	1	2

Term	D-3	D-6
negosiasi	0	8
lebih	0	4
dasar	0	3
ikut	0	2
kantor	0	2
rapat	0	2
orang	0	2
manusia	0	2
teknik	0	2
kalangan	0	2
bangun	0	2

D-3*D-6 0.00

|D-3| 0

|D-6| 11

$\cos\theta =$ D-3.D-6/(|D-3||D6|)
= Δ

Term	D-4	D-5
ruang	0	5
bunyi	0	5
gema	0	2
cegah	0	2
gedung	0	2
serap	0	2
keras	0	2
langit	0	2
dinding	0	2
musik	7	7
budaya	4	4
lampung	4	4
festival	3	3

D-4*D-5 108.00

|D-4| 10.34408

|D-5| 13.85641

$\cos\theta =$ D-4.D-5/(|D-4||D5|)
= 0,75

Lampiran 4. (Lanjutan)

Term	D-4	D-5
ada	2	2
hingga	2	2
tradisional	2	2
daerah	2	2
bahan	0	2
akustik	1	2

Term	D-4	D-6
negosiasi	0	8
lebih	0	4
dasar	0	3
ikut	0	2
kantor	0	2
rapat	0	2
orang	0	2
manusia	0	2
teknik	0	2
kalangan	0	2
bangun	0	2

D-4*D-6 0.00

|D-4| 0

|D-6| 11

$\cos\theta =$ D-4.D-6/(|D-4||D6|)
= Δ

Term	D-5	D-6
ruang	5	0
bunyi	5	0
gema	2	0
cegah	2	0
gedung	2	0
serap	2	0
keras	2	0
langit	2	0
dinding	2	0
musik	7	0
budaya	4	0
lampung	4	0
festival	3	0
ada	2	0
hingga	2	0
tradisional	2	0
daerah	2	0
bahan	2	0
akustik	2	0

D-5*D-6 0.00

|D-5| 13.85641

|D-6| 11

$\cos\theta =$ D-5.D-6/(|D-5||D6|)
= 0

Lampiran 4. (Lanjutan)

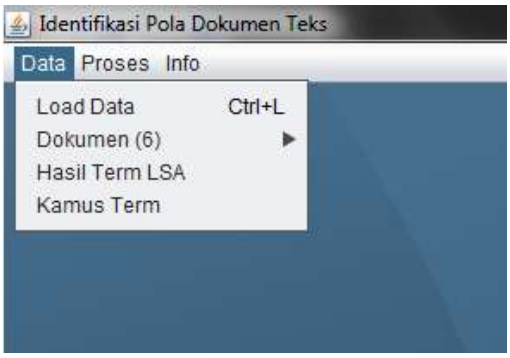
Term	D-5	D-6
negosiasi	0	8
lebih	0	4
dasar	0	3
ikut	0	2
kantor	0	2
rapat	0	2
orang	0	2
manusia	0	2
teknik	0	2
kalangan	0	2
bangun	0	2

Lampiran 5. Buku Manual Program Bab 6

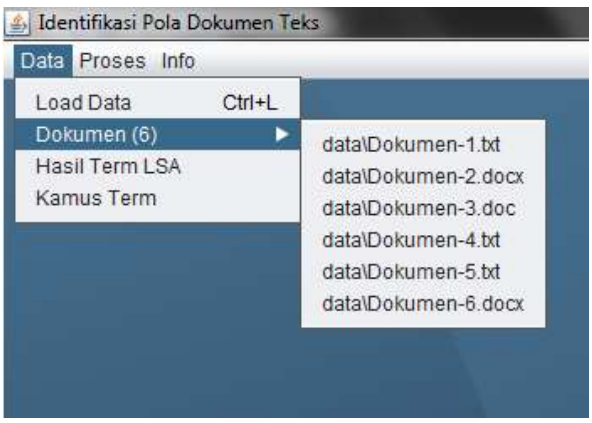
Untuk menjalankan program Bab 6 diperlukan langkah-langkah sebagai brikut:

1. Buka (klik Program Bab 6), keluar sub menu,

2. Klik Data pada sub menu (untuk mengambil data berupa dokumen), lalu tekan Loading Data akan muncul sub menu sebagai berikut:

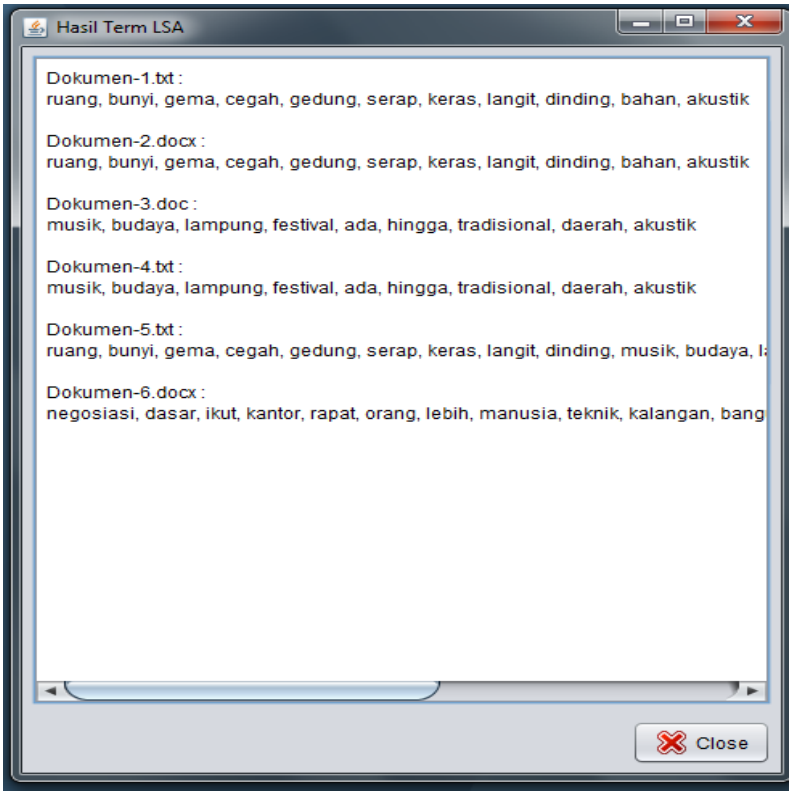


3. Klik Dokumen untuk melihat isi masing-masing dokumen

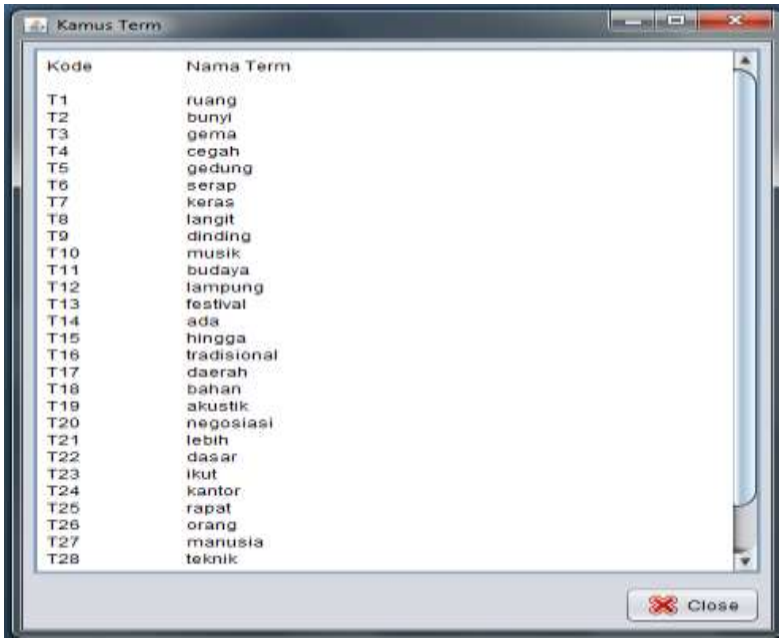


Lampiran 5. (lanjutan)

4. Klik Hasil *Term* LSA untuk melihat *term* untuk masing-masing dokumen.

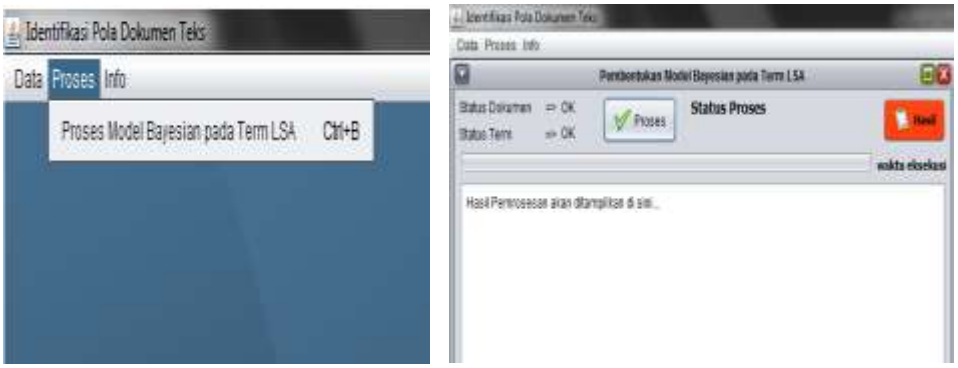


5. Klik Kamus Term untuk melihat kumpulan term seluruh dokumen yang telah diberi identitas untuk masing-masing term.

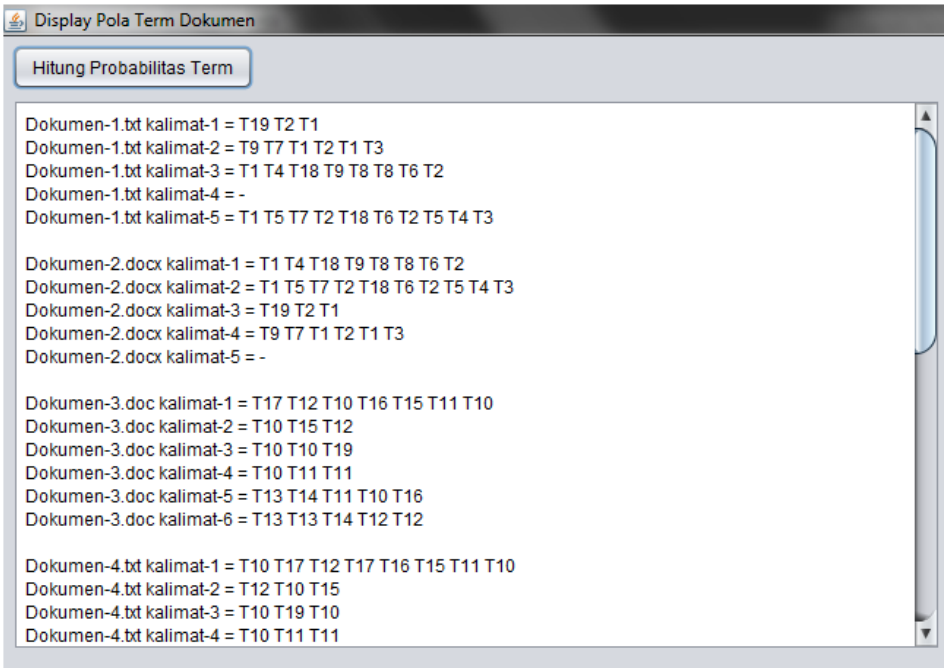


Lampiran 5. (lanjutan)

6. Klik Proses, untuk memproses perhitungan Bab 6.



7. Klik Yes untuk menampilkan order term untuk masing-masing dokumen.



8. Klik Hitung Probabilitas *Term* untuk menuju sub menu perhitungan probabilitas munculnya term, perhitungan *likelihood* masing-masing kalimat disetiap dokumen, menghitung *likelihood* masing-masing dokumen, menghitung rasio *likelihood* antar dokumen dan menampilkan perhitungan rasio *likelihood* untuk setiap pasangan dokumen dari enam dokumen.

Lampiran 5. (lanjutan)

a. Perhitungan *likelihood* masing-masing kalimat disetiap dokumen:

Likelihood Per Kalimat

Dokumen	Kalimat	Prob Term-1	Prob Term-2	Prob Term-3	Likelihood
Dokumen-1.txt	kalimat-1	0.0833	0.0833	0.0833	0.0006
Dokumen-1.txt	kalimat-2	0.0833	0.0833	0.0833	0.0006
Dokumen-1.txt	kalimat-3	0.1667	0.0833	0.0833	0.0012
Dokumen-1.txt	kalimat-5	0.1667	0.0833	0.0833	0.0012
Dokumen-2.docx	kalimat-1	0.1667	0.0833	0.0833	0.0012
Dokumen-2.docx	kalimat-2	0.1667	0.0833	0.0833	0.0012
Dokumen-2.docx	kalimat-3	0.0833	0.0833	0.0833	0.0006
Dokumen-2.docx	kalimat-4	0.0833	0.0833	0.0833	0.0006
Dokumen-3.doc	kalimat-1	0.0278	0.0278	0.0278	2.1433E-05
Dokumen-3.doc	kalimat-2	0.2500	0.0278	0.0278	0.0002
Dokumen-3.doc	kalimat-3	0.2500	0.0278	0.0278	0.0002
Dokumen-3.doc	kalimat-4	0.2500	0.0833	0.0833	0.0017
Dokumen-3.doc	kalimat-5	0.0556	0.0278	0.0278	4.2867E-05

b. Perhitungan *likelihood* disetiap dokumen:

Likelihood Per Dokumen

Dokumen	Likelihood
Dokumen-1.txt	4.4863E-13
Dokumen-2.docx	4.4863E-13
Dokumen-3.doc	2.5444E-24
Dokumen-4.txt	2.0844E-20
Dokumen-5.txt	9.3510E-33
Dokumen-6.docx	9.6952E-29

c. Perhitungan rasio *likelihood* antar dokumen:

Faktor Bayes antar Dokumen

Pembilang	Penyebut	Nilai Rasio Likelihood
Dokumen-1.txt	Dokumen-2.docx	1.0
Dokumen-1.txt	Dokumen-3.doc	1.7632E+11
Dokumen-1.txt	Dokumen-4.txt	2.1523E+07
Dokumen-1.txt	Dokumen-5.txt	4.7976E+19
Dokumen-1.txt	Dokumen-6.docx	4.6273E+15
Dokumen-2.docx	Dokumen-3.doc	1.7632E+11
Dokumen-2.docx	Dokumen-4.txt	2.1523E+07
Dokumen-2.docx	Dokumen-5.txt	4.7976E+19
Dokumen-2.docx	Dokumen-6.docx	4.6273E+15
Dokumen-3.doc	Dokumen-4.txt	8192.0
Dokumen-3.doc	Dokumen-5.txt	2.7210E+08
Dokumen-3.doc	Dokumen-6.docx	26244.0000
Dokumen-4.txt	Dokumen-5.txt	2.2290E+12

Lampiran 5. (lanjutan)

d. Perhitungan perhitungan rasio *likelihood* untuk setiap pasangan dokumen dari enam dokumen:

Faktor Bayes 6 Dokumen

	Dokumen-1	Dokumen-2	Dokumen-3	Dokumen-4	Dokumen-5	Dokumen-6
Dokumen-1		1.0	1.7632E+11	2.1523E+07	4.7976E+19	4.6273E+15
Dokumen-2			1.7632E+11	2.1523E+07	4.7976E+19	4.6273E+15
Dokumen-3				8192.0	2.7210E+08	26244.0000
Dokumen-4					2.2290E+12	2.1499E+08
Dokumen-5						10368.0

9. Klik info, menuju sub menu info tentang program Bab 6 sebagai berikut:

DAFTAR RIWAYAT HIDUP

I. Data Pribadi

Nama	:	Soehardjoepri
NRP	:	1310301002
Tempat/ Tgl.Lahir	:	Jember, 04 Mei 1962
Agama	:	Islam
Instansi Satuan Kerja	:	Jurusan Matematika FMIPA-ITS
NIP	:	19620504.198701.1.001
Alamat Rumah	:	Raya Wiguna Utara 44, Surabaya HP. 08165413862
E-mail	:	djoepri.its@gmail.com

II. Latar Belakang Pendidikan

No	Nama Pendidikan	Jurusan Bidang	Tahun Lulus
1.	SD Negeri 2, Kencong, Jember	-	1974
2,	SMPN 1, Kencong, Jember	-	1977
3,	SMAN 1, Jember	IPA	1981
4.	S1 FMIPA-ITS, Surabaya	Matematika	1986
5.	S2 FMIPA-UGM, Yogjakarta	Matematika	1998

III. PUBLIKASI

Publikasi yang dilakukan dalam bentuk Jurnal dan Seminar selama masa studi

Jurnal Internasional:

- Soehardjoepri, Iriawan, N., Ulama, B.S.S. dan Irhamah (2015), "Identifying Text Document Pattern For Two Terms Appearances VIA Latent Semantic Analysis (LSA) Method And Term Distance Between Two Documents", *Journal of Theoretical and Applied Information Technology*, Vol. 79, No. 2, hal. 322-329.
- Soehardjoepri, Iriawan, N., Ulama, B.S.S. dan Irhamah (2016), "On The Identification Of The Structural Pattern Of Terms Occurrence In a Document Using Bayesian Network", *Journal of Theoretical and Applied Information Technology*, Vol. 92, No. 2, hal. 253-264.

Seminar Internasional:

- Soehardjoepri, Iriawan N., Ulama, B.S.S. dan Irhamah (2013), "On the Text Documents Pattern Recognition Using Latent Semantic Analysis and Kolmogorov-Smirnov Test", *Proceedings South East Asian Conference on Mathematics and Its Applications*, Department of Mathematics, FMIPA-ITS, Surabaya, hal. AM-24.